This redlined draft, generated by CompareRite (TM) - The Instant Redliner, shows the differences between -
original document : C:\WINDOWS\DESKTOP\APPLN_ORIG.WPD
and revised document: C:\WINDOWS\DESKTOP\APPLN_REV.WPD

CompareRite found 478 change(s) in the text

Deletions appear as Normal text surrounded by {}
Additions appear as Bold-Underline text

METHOD AND MEANS FOR SORTING AND IDENTIFYING BIOLOGICAL INFORMATION

This application is a continuation-in-part of U.S. Patent Application Serial No. 07/201,358, filed May 26, 1988, which is hereby incorporated by reference herein, which application is a continuation of U.S. Patent Application No. 06/770,390, filed August 28, 1985, now abandoned.

FIELD OF THE INVENTION

This application describes discrete populations of oligopeptides of random sequences, polypeptides comprising those oligopeptides, oligonucleotides encoding those oligopeptides and recombinant vectors comprising those oligonucleotide sequences. The population of oligopeptides represents the universe of peptide epitopes. Also disclosed are discrete populations of antibodies (or hybridomas) capable of binding to the populations of oligopeptides. The disclosure of the present application relates to the identification and characterization of peptide epitopes, or recognition sites, of antibodies. More particularly, the determination of the linear amino acid sequence recognized by the antibody and of a nucleic acid sequence encoding that amino acid sequence are enabled by the disclosure herein.

BACKGROUND OF THE INVENTION

This invention relates to the characterization and identification of the recognition sites of antibodies. More particularly, this invention involves the determination of the specific amino acid sequence recognized by an antibody, and of the nucleic acid sequence encoding that amino acid sequence. The clonal selection theory of Burnet, which explains the general basis of antibody production, has gained

virtually complete acceptance. Burnet, M. (1961) Sci. {Am.} **Ara.** 204:58; Jerne, N.K. (1976) Harvey Lecture 70:93. The theory is based on several premises: (1) as individual cells, i.e., lymphocytes, in the immune system differentiate, each becomes capable of producing only one species of antibody molecule; (2) the entire spectrum of possible antibody-producing cells is present within the lymphoid tissues prior to stimulation by any antigen; that is, the step in which each lymphocyte becomes specified to produce only one type of antibody molecule occurs in the absence of a potential antigen for that antibody; and (3) lymphocytes capable of producing an antibody specific to a particular antigen are induced, by the presence of that antigen, to proliferate and to produce large quantities of the antibody. An enormous range of genetically unique lymphoid cells is present in the lymphoid organs, e.g., the spleen, of each mammal. The spleen can be considered a library of cells, each of which can manufacture a unique antibody, and the library is so large that for any {arbitrary} **particular** antigen, at least one lymph cell exists within the library that is capable of recognizing the antigen and producing antibodies specific to the antigen.

Heretofore, the production of an antibody that will recognize an antigen of interest has required the antigenic stimulation of a laboratory animal. Typically, the antigen is injected into a laboratory animal, and, after a suitable incubation period, a second injection is given. The spleen cells of the animal are then harvested and fused to myeloma cells. When fused to a spleen cell, the myeloma cell confers to the spleen cell its ability to grow in culture. Surviving colonies of fused cells, i.e., hybridomas, are then screened to identify clones that produce antibodies that specifically recognize the antigen. This procedure must be repeated each time it is desired to produce an antibody to a particular antigen. For each antigen of interest, it is necessary to (1) antigenically stimulate an animal, (2) remove its spleen and hybridize the spleen cells with myeloma cells, and (3) dilute, culture, and screen clones for specific antibody

production. Though antibodies that recognize the antigen are produced, this technique does not identify the epitope, i.e., the specific site on the antigen that an antibody recognizes; and one cannot direct the development of antibodies specific to a particular predetermined site or region of the **antigen.** Also, hybridoma techniques are not effective in the direct development of monoclonal antibodies that recognize haptens, i.e., molecules that contain **constitute** antibody recognition sites, but which do not elicit an antigenic reaction when injected without a carrier into a laboratory animal. Since antigenic stimulation and antibody production are potentially hazardous to the host, the use of human hosts has been precluded in the development of monoclonal antibodies.

The universe of antibody binding specificities may be open or closed. If the universe of antibody binding specificities is closed, then the following basic tenets apply:

a) **one can design and prepare any given epitope and isolate any antibody (for example, a monoclonal antibody produced by a member of a random set of hybridomas) from a universe of antibodies without having first immunized an experimental animal with an antigen containing that epitope. A self-addressing sorting scheme can be used to screen to identify the proper paired correspondence between antibody and epitope;**

b) **the universe of epitopes can be specified in at least a theoretical fashion, and in principle, can be synthesized; and**

c) **one can independently isolate and identify an antibody-producing hybridoma with the same epitopic specificity as one previously isolated and identified. Such a repeated isolation occurs in a "second hit" experiment, and can be used to estimate the effective**

size of the universe of antibody specificities. Such an approach is similar in logic to defining a complementation group in genetics.

Even if the universe of epitopes is large, if it is closed, it can be defined by rules, algorithms or iterative analyses.

In the alternative, if the universe of antibody specificities is open, the following principles apply:

a) one cannot isolate an antibody specific for an epitope without prior immunization with an antigen containing that epitope;

b) the universe of epitopes cannot be specified or synthesized; and

c) one should not be able to independently isolate more than one antibody with the same target specificity.

The binding domain of a monoclonal antibody specific to a malaria virus surface protein has been identified as being no larger than 40 amino acids long. Cochrane, A. H. et al{,}. Proc. Natl. Acad. Sci. U.S.A. 79:5651 (1982), inserted a 340 base pair sequence from a Plasmodium knowlesi gene into the {pBR 322} **pBR322** vector. The engineered vector produced in E. coli a beta-lactamase fusion polypeptide that reacted with a monoclonal antibody specific {to} **for** a P. knowlesi {circumsporozoite or CS} **circumsporozooite (CS)** protein. This finding indicated that the binding domain of the monoclonal antibody was limited to a region of the CS protein encoded by the inserted sequence, or approximately 110 amino acids. Lupski, {J.R} **J. R**. et al., Science 220:1285 (1983), used the same system and, employing transposition mapping techniques, further localized the binding domain to a 40-amino{,} acid region of the CS protein.

Green, N. et al., published PCT application {84/00 687} **84/00687**, produced antibodies by {innoculating} **inoculating** laboratory animals with synthetic peptides. Antibodies produced in response to peptides having a length of 8 to 40 amino acid

residues and corresponding to sequences in an influenza virus protein were cross-reactive with the virus in vitro.

Dame, {J.B} **J. B**. et al., Science 225:593 (1984), sequenced the CS gene of Plasmodium falciparum and discovered 41 tandem repeats of a tetrapeptide, with some minor {variation. Using} **variations. Using problem** synthetic peptides of 4, 7, 11, and 15 amino acid residues of the predominant repeating amino acid sequence, Dame **et al.** then conducted competitive binding assays to determine what length of peptide would inhibit the binding of the CS protein with a monoclonal antibody specific to that protein. Dame **et al.** found that the synthetic 4 amino acid sequence did not significantly inhibit binding, but the 7, 11 and 15 amino acid sequences did inhibit binding. These results suggest that this monoclonal antibody to the CS protein recognizes a 5 to 7 amino acid sequence {containing} **comprising** the repeating tetrapeptide.

The known crystal structures of the Fab fragment and lysozyme show that there are two contact points on the lysozyme molecule for the antibody combining site, and each contact point spans over about five amino acids. Earlier work on antibody binding to carbohydrate antigens and glycosidase cleavage protection experiments show that 5-6 sugar residues are protected from glycosidase cleavage. Studies with antibody binding to haptens also suggests that antibody sites are small. Peptide competition experiments, also called epitope mapping experiments, show that oligopeptides 4 to 5 amino acids in length can specifically compete for antibody binding.

In addition, linear sequences which differ in only one amino acid, can compete for antibody binding with varying degrees of specificity (see, e.g., Geysen et al. (1986) in Synthetic Peptides as Antigens; Ciba Foundation Symposium 119, Porter and J. Wheelan, Eds. (New York, Wiley) pp. 130-149).

While five amino acids is a representative length of peptide sequence which can bind with differential specificity to an antibody, five amino acid residues is not necessarily the size of an immunogenic peptide. Generally, when an oligopeptide is the desired immunogen, it is first conjugated to a larger carrier molecule. The actual operational relationship between the immunizing entity and the binding entity can only be resolved when an in vitro immunization-dependent antibody synthesis system is developed.

SUMMARY OF THE INVENTION {Summary of the Invention}

In one aspect the invention features a **discrete** population of oligonucleotides, each {containing between 1 and about 50 tandem sequences of} **comprising** the same length of from about 4 to about 12 nucleic acid **coding** triplets **in random order**. Each oligonucleotide encodes {for} a corresponding oligopeptide of **from** about 4 to about 12 L-amino acid residues, and the entire population represents at least about 10% of all oligopeptide sequences of the selected length. In preferred embodiments, each member of the oligonucleotide population has a single copy of the **random** sequence of nucleotide triplets, the oligonucleotide sequence has between {5} **4** and 7 triplets, and the

oligonucleotide population can be generated by random shearing of mammalian genetic material or is chemically synthesized from the component {nucleic acids.} **nucleotides.**

It is particularly preferred that each oligonucleotide sequence comprises five coding triplets. The oligonucleotide population may also be composed of members, each of which contains the same number of tandem repeats of each peptide coding sequence, where the number of tandem repeats is from two to about fifty. It is particularly preferred that the oligonucleotide population be

sufficiently redundant so that each of all possible encoded oligopeptide sequences is present at least 10 times on average.

In a second aspect the invention features a **discrete** population of oligopeptides {containing between 1 and about 50 tandem sequences} **each of random amino acid sequence** of the same length, of about 4 to about 12 **L-**amino acid residues, and the population makes up at least 10% of all peptide sequences of the predetermined length. In preferred embodiments each member of the population has a single copy of the peptide sequence, the oligopeptide sequence has between {5} **4** and 7 L-amino acid residues, and the population {is} **can be** generated by shearing of proteins {or is chemically synthesized}**, by chemical synthesis** from the component L-amino acids**, or by the translation of the oligonucleotides of random coding sequences.**

It is particularly preferred that there be five amino acid residues in each oligopeptide. It is particularly preferred that the population of oligopeptides is sufficiently large so that each sequence is represented at least 10 times on average. The peptide population can also be composed of member peptides, each of which contains the same number of tandem repeats of the amino acid sequence, where the number of repeats is from two to about fifty.

In a third aspect, the invention features a **discrete recombinant** vector population of substantially identical autonomously replicating nucleic acid sequences including a structural gene and a population of oligonucleotide inserts **therein, each insert** containing {between 1 and about 50 tandem sequences of} a uniform length selected from between about 4 to about 12 nucleic acid {triplets, each} **coding triplets, preferably between 4 and 7, and most preferably five. Each** insert is recombinantly inserted **in frame** into the structural gene of one of the nucleotide sequences, and **preferably** the oligonucleotide population encodes {for at least about

10% of} all oligopeptide sequences of the predetermined length. **Preferably the recombinant vector population is redundant, i.e., contains a sufficient number of random oligonucleotide members so that all possible members are represented at least once. It is particularly preferred that the population is sufficiently redundant so that the population contains at least 10 copies of oligonucleotides encoding each possible peptide sequence, on average.** In preferred embodiments each member of the insert population has a single copy of the sequence of nucleotide triplets, and the insert has {between 5 and 7} **coding** triplets; the replicating sequence can be a plasmid such as pBR322 {a virus such as lambda-gt 11 or vaccinia, or a hilamentous bacteri}**, a virus such as λgt11 or vaccinia, or a filamentous bacteriophage, such as fl, fd or M13. The recombinant vector population can also be made up of individual vectors each containing the same number tandem repeats of an oligonucleotide sequence as defined above. The number of tandem repeats can be from two to about fifty in number.**

The recombinant vector population can also be made up of individual vectors each containing the same number tandem repeats of an oligonucleotide sequence as defined above. The number of tandem repeats can be from two to about fifty in number.

In a fourth aspect, the invention features a **discrete** heterogeneous population of **antibodies comprising member** antibodies capable of binding to substantially all members of {an} **discrete** oligopeptide population featured in the second aspect of the invention, above.

In a fifth aspect, the invention features a **discrete** population of binding pairs that includes {a} **the discrete** population of peptide sequences all of the same length {of} **selected from** about 4 to about 12 L-amino acid residues and {a} **the** heterogeneous population of antibodies capable of binding to substantially all the

peptide sequences, where substantially every member of the peptide population is bound to a corresponding antibody.

In a sixth aspect, the invention features a matrix including a **discrete** population of **random** peptide sequences and a heterogeneous population of antibodies.

In a seventh aspect, the invention features a method for constructing a matrix including the steps of (1) obtaining a population of **peptides or** polypeptides **comprising peptides as described above,** having a uniform length of between about 4 and about 12 **L**-amino acid residues **of random sequence** and including at least about 10% of all peptide sequences of the predetermined length; (2) obtaining a **discrete** heterogeneous population of antibodies capable of binding to substantially every member of the polypeptide population; and (3) contacting the antibodies with the antigens for a sufficient amount of time and under appropriate conditions so that binding occurs. **Preferably, the peptide length is 4 to 7 amino acids, and most preferably, 5 amino acids.** In preferred embodiments**:** each of the {peptide sequences and }**peptides and each of the** antibodies is isolated and each {peptide sequence} is contacted individually with each of the antibodies until at least one peptide {sequence-}antibody binding pair is identified; the peptides can be immobilized on an appropriate substrate and the antibodies can be labeled; the antibodies can be immobilized and the **peptide** sequences can be labeled; or the peptide sequences can be excised from the polypeptides.

It is preferred in all of the foregoing aspects of the invention that the populations be sufficiently large so as to contain all theoretical members of the population, and it is particularly preferred that each population of the invention is sufficiently redundant so that it is statistically unlikely that sampling for a particular member will fail, as is understood in the art.

The invention provides an efficient and convenient means for the **identification and** production of monoclonal antibodies to any specific region of any antigen or hapten of interest. Monoclonal antibody production, according to the invention, does not require antigenic stimulation of a host animal. **This is a critical concept of the present invention. Such antigenic stimulation can be employed to increase the frequency for cognate hybridoma formation, but there will be a member of an antibody population (of a sufficiently large number of members) which will recognize the particular epitope even in the absence of such stimulation.**

The invention involves the antibody binding properties of a test species, **e.g., a peptide,** but is totally independent of the ability of the test species to induce an antigenic response in vivo. The invention permits the identification of the specific peptide sequence on a protein that is recognized by an antibody**, i.e., the epitope**. The specificity of antibodies recognizing distinct sequences**, or epitopes,** on the same antigen can be differentiated. In addition, the invention permits the characterization and the localization on a chromosome of the nucleotide sequence encoding {for} the amino acid sequence recognized by an antibody.

Using conventional monoclonal techniques, one can produce antibodies that might react, for example, with an undetermined site on a particular Plasmodium {circumsporozoite} **circumsporozooite** protein or a particular influenza virus. Using the present invention, one can identify all the epitopes on that molecule or organism and obtain {an antibody that recognizes} **antibodies recognizing** each of these epitopes. {An epitope is a specific site on the surface of an antigen that is recognized by an antibody.} By judiciously combining a number of distinct antibodies, each of which recognizes a different epitope on the surface of a particular antigen, a material with any desired degree of specificity can be obtained. Also using the invention, one

can identify **epitopic** sequences that are common to, e.g., the {circumsporozoite} **circumsporozooite** proteins of several Plasmodium species or **common** to several strains of influenza, and screen for antibodies recognizing these common sequences, thereby identifying a single set of antibodies, each of which is effective against a broad range of malarial or influenza infections.

Certain viruses, such as the LAV or HTLV-III virus, contain on their surfaces both highly mutable regions and constant regions. The {viruses'} **viruses'** ability to alter their surface characteristics has hampered the development, through standard monoclonal techniques, of antibodies to these viruses. Any antibody that recognizes a mutable region of a virus would become ineffective as the virus mutated {and a strain developed} **to produce strains** having {an} altered {configuration} **configurations** in the region recognized by the antibody. Once the constant regions of a virus have been identified and characterized, the invention permits the identification and production of antibodies that recognize these constant regions, even if the peptide sequences comprising these constant regions would not themselves elicit an immunogenic response in vivo. Such antibodies would be effective against various **mutated** strains of the virus.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments and from the claims.

It is believed that an epitope has limited dimensions of between about 30 and 50 angstroms. An antibody that recognizes a specific peptide sequence or configuration {of} **or** carbohydrates on the surface of an antigen will recognize that same configuration if it is duplicated or closely approximated on a different antigen. This phenomenon underlies the cross-reactivity sometimes encountered with monoclonal antibodies.

The size of the antibody recognition site corresponds to a peptide sequence in the range of between about 4 and {about 12} 7 amino acid residues **with the majority of recognition sites spanning about 4 to 6 amino acids**. Mammalian proteins and polypeptides are composed almost exclusively of the twenty naturally occurring amino acids, i.e., glycine and the L-isomers of alanine, valine, leucine, isoleucine, proline, phenylalanine, tyrosine, tryptophan, serine, threonine, aspartic acid, glutamic acid, asparagine, glutamine, cysteine, methionine, histidine, lysine, and arginine. There are about three million different possible sequences of the twenty amino acid residues taken five at a time, and about sixty million if the amino acid residues are taken six at a time. This finite number of peptide sequences {may represent} **represents** the full range of possible antibody recognition sites **which can be represented or mimicked by linear peptide epitopes**. Production and maintenance of a representative sample of the full range of antibodies and of a representative sample of the peptide sequences of the appropriate length provides the means (1) to screen any antibody of interest in order to determine the precise **epitopic** peptide sequence it binds to and (2) to screen any protein in order to find an antibody specific to that protein.

The present invention identifies **epitopic** (antibody binding) **(antibody binding)** sites that comprise a primary peptide sequence {or}**. The identified linear epitope may mimic a discontinuous peptide epitope or a non-peptide epitope**, e.g., a carbohydrate sequence that {is} **can be** closely approximated by a peptide sequence **with respect to antibody recognition.**

In view of these considerations {Notwithstanding these beliefs}, the invention provides the means and methods for the identification and characterization of **peptide** epitopes, and of the antibodies that bind to them.

DETAILED DESCRIPTION OF THE INVENTION

Antibody Production

According to the clonal selection theory, an unchallenged mammalian host has the capacity to produce antibodies to a vast array of foreign antigens. The presence of an antigen triggers the proliferation of those lymphocytes already present having the ability to produce antibodies {to the} **specific for that** antigen. Since there is a finite number of **linear** peptide sequences of the length that is recognized by antibodies, it can be expected that each mammal has the capability to produce antibodies that will recognize most, if not all of these sequences. Thus, the spleen of a mouse or another laboratory animal can serve as an appropriate source for a full range of antibodies. The spleen can be harvested from a laboratory animal, and, using standard techniques, the individual cells are fused to myeloma cells and hybridoma strains are developed.

Depending on the desired characteristics of the resulting hybridoma population, either antigenically stimulated animals can be used, or animals that have not been specifically challenged with the antigenic material of interest can be used.

If antigenically stimulated animals are used, then a higher proportion of the resulting hybridomas will produce antibodies specific to the antigen used. If, on the other hand, unchallenged animals are used, then it can be expected that the antibodies retrieved from the resulting population of hybridomas will represent a broader range of the antibodies that the animals are capable of producing. The **predominant** antibodies produced by a mature animal raised under standard laboratory conditions will reflect and be limited by its individual exposure history. If spleens are harvested from several **(at least about 10)** unchallenged mature animals and combined together, and the spleen cells fused to myeloma cells, then the resulting **discrete** population of hybridomas will produce a more complete range of antibodies then would hybridomas

from any single individual. Antibodies produced by the hybridomas derived from the spleen cells of mature **animals that were raised aseptically or from fetal or neonatal** animals that were raised aseptically or from fetal or neonatal animals will not reflect any exposure history and can be expected to represent a random sample of the full range of antibodies that the animals are capable of producing.

Since this procedure does not require antigenic stimulation of {the} donor {animal} **animals** before harvesting the {spleen} **spleens**, it is now possible to develop antibodies derived from human cells. Normal spleen cells can be collected from one or a number of human donors and the harvested cells fused to myeloma cells and cultured as described above. Alternatively, a library of human antibodies can be developed over time by obtaining cell cultures from, e.g., a large number of myeloma patients, each patient having a distinctive tumor.

It is now possible to use a recombinant library to generate the universe of antibody binding specificities instead of a hybridoma library. Huse et al. (1989) Science 246:1275-1281, describes the generation of a large combinational library of mouse Fab fragments. Alting-Mees et al. (1990) Strategies in Molecular Biology 3:1-2,9 describes bacteriophage (λ) expression libraries for antibody production.

Production of Peptide Sequences {Production of peptide sequences}

Numerous methods are available for the production of the desired population of peptide sequences. For certain embodiments of the invention these peptide sequences can be produced directly either by randomly shearing proteins and then recovering by electrophoresis the peptide sequences of the appropriate length, or by synthesizing the desired **random** peptide sequences from {their} **the** component amino acids.

Alternatively, these peptides can be produced through genetic engineering techniques. Peptides produced according to this general method can be termed coded peptides. A population of nucleotide sequences {is first obtained} of the correct length to encode {for} **random** peptide sequences of the desired length **is generated**. This can be accomplished either by random cleavage of biological genetic material followed by electrophoresis to recover those nucleotide sequences that were **cut or** sheared to the desired length, or by **chemical** synthesis from the component {nucleic acids} **nucleotides or codons**.

Depending on the desired characteristics of the resulting population of nucleotide sequences and ultimately, of the peptide sequences to be produced, different techniques are used to obtain the population of nucleotides. If a random population of nucleotide sequences is desired, then the nucleotides can be synthesized by adding the four {nucleic acids} **nucleotides** with equal frequency at each position of the growing nucleotide chains. If it is desired that the synthesized nucleotide triplets more closely reflect the distribution of naturally occurring triplets, then the frequency of each {nucleic acid} **nucleotide** employed at the first, second, or third position of each triplet can be manipulated to approximate the frequencies at which each {nucleic acid} **nucleotide** residue appears at each position in nature, as suggested in Crick F.H.C. et al., {Origin} **Origin** of Life, 7:389-397 (1976). Any of several sources of genetic material can be selected to obtain by shearing nucleotide sequences of the desired length, e.g., cellular DNA or cDNA. cDNA} **cDNA. CDNA**, of course, would provide a {tighter} **closer** representation of the naturally occurring coding sequences. **Alternatively, chemically synthesized oligonucleotides of tandem sequence may be used.**

When the desired population of nucleotide sequences has been obtained, the population can then be treated to facilitate the insertion of each sequence into a vector

and to facilitate the subsequent recovery of the desired peptide sequence from the culture of host cells incorporating the engineered vector. For example, using known techniques, AUG sequences can be ligated to each end of each member of the population of {nucleotides} **nucleotide sequences**. When each nucleotide **sequence** is translated, the desired peptide sequence will be flanked by methionine residues. The translated protein can then be treated with cyanogen bromide, which cleaves peptides at methionine sites, to excise the desired peptide sequence from the protein. The cleavage product can then be purified by {;} electrophoresis. {Alternatively} **Preferably**, a restriction endonuclease recognition sequence can be ligated to each end of each member of the population of nucleotide **sequences** and then the population of {nucleotides,} **nucleotide sequence** can be treated with the endonuclease recognizing the ligated sequence to produce "sticky ends" which facilitate the insertion of the nucleotide **sequence** at the restriction site in a vector recognized by the endonuclease. **When the population of nucleotide sequences is chemically synthesized, flanking restriction sites may be designed into the oligonucleotide nucleotide sequence, as understood in the art.**

Each nucleotide sequence {Each nucleotide} is then inserted into an appropriate **vector. The ratio of nucleotide sequences to vectors can be** controlled to ensure that**, on the average,** no more than one nucleotide sequence is inserted into any vector. The nucleotide sequence must be inserted at a location in the vector where it will be translated in phase when the vector is transferred into an appropriate host cell, and where it will not interfere with the replication of the vector under the experimental conditions employed{. The}**, i.e., the** nucleotide sequence must be inserted into a **nonessential** region of the vector. Pieczenik, U.S. Patent **Nos.** 4,359,535**, and 4,528,266** hereby incorporated by reference, {discloses} **disclose** a method for inserting foreign DNA into a non-essential region of a vector.

Smith (1985) Science 228:1315-1317 describes the insertion of heterologous coding sequences into the unique BamHI within the minor coat protein (PIII) gene (gene III) of λ and immunological screening for recombinant phage expressing the heterologous coding sequence. Parmley and Smith (1988) Gene 73:305-318 describe an λ derivative which allows for the insertion of heterologous coding sequences at an engineered cloning site, allowing for the expression of a heterologous coding sequence near the mature N-terminus of pIII. Immunoaffinity purification can be used to purify recombinant phage expressing a desired epitopic sequences).

The nucleotide sequence is advantageously inserted in such a way that the peptide sequence encoded by the nucleotide sequence is expressed on the outside surface of the {vector} **bacteriophage or the host cells with plasmids containing the nucleotide sequence**. To prepare inserts having these characteristics, {an appropriate} **a** vector, e.g., a phage or plasmid, {is first selected. The vector is then randomly cleaved }**with an appropriate cloning site, is first selected.**

A suitable position for a cloning site may be determined empirically by performing an experiment to identify an insertion site in a structural gene which will allow expression of an inserted oligonucleotide coding sequence, and which will result in the expression of the encoded oligopeptide as an epitope within or at one end of a structural gene product such that recognition of the epitope in the recombinant virus or genetically modified host cell or protein is possible. That oligopeptide sequence can be detected using an antibody specific for an epitope of that sequence (or specific for an epitope mimicked by the conformation of that sequence).

The vector can then be cleaved at random sites according to the method disclosed in {Pieczenik,} U.S. Patent 4,359,535{,} **and 4,528,266** to yield a

population of linear DNA molecules having circularly permuted sequences**, where the breakpoint in the circular molecule is at a random location in each molecule**.

After the cleavage steps, a synthetic oligonucleotide linker bearing a unique nucleotide sequence not. present on the original unmodified vector can be attached to both ends of each linearized vector by blunt end ligation. The random {linears} **linear DNA molecules** can then be treated with the restriction endonuclease specific to the attached sequences, to generate cohesive ends.

All such recombinant vectors which allow immunologic detection of the encoded oligopeptide express that epitope in a context-insensitive fashion. For the purposes of this invention, context-insensitive means that the milieu in which the oligopeptide is expressed does not prevent recognition by the cognate antibody. The actual insertion site on the vector can be determined by sequence analysis, as understood in the art, and that site can be modified to contain an appropriate cloning site. As understood in the art, the insertion and immunological detection should be repeated to confirm functionality in context-insensitive expression of an epitopic sequence. such an engineered vector can be used in the practice of the invention. The immunological detection of an inserted oligonucleotide sequence encoding a context-insensitive epitope is to be called a "topological mapping" of the surface of the vector. The topological mapping of a vector allows the optimum design of an expression vector.

DNA sequences {DNA} encoding a gene product, e.g., human hemoglobin,{,not } **where these sequences are not naturally** present in the vector, {is} **can be cleaved by any method known to the art and** fractionated to the desired size, e.g., fifteen nucleotides long, and the nucleotide sequences ligated to the same type of linker used with the random linears. The fractionated nucleotide sequences are then inserted into the random linears, and the modified vectors are transferred into

appropriate host cells. The host cells are diluted, plated, and the individual colonies **(or plaques)** grown up. **On** replica plates, the colonies **(or plaques)** are screened with a monoclonal or polyclonal antibody specific to the gene product. **A suitable control to insure that selected colonies or plaques express epitopes of the desired specificity is the host cell into which unmodified vector has been introduced, as understood by the skilled artisan.**

A positive reaction with the antibody identifies a colony wherein the inserted nucleotide sequence is translated in phase, and the encoded peptide sequence is on the outside surface of the polypeptide or protein, **or otherwise** accessible to the antibody screening assay. If a monoclonal antibody is employed in the screening step, then this procedure will identify only those colonies where the specific peptide sequence comprising the site recognized by that antibody is inserted on the outside surface of the polypeptide or protein **unless appropriate pretreatment has been carried out**. If a polyclonal antibody is employed, or a mixture of several monoclonals, then any colony {containing on the outside surface of the}**, virus,** polypeptide or protein {any peptide sequence insert comprising a recognition site of the foreign gene product} **expressing a cognate epitope in a manner accessible for antibody binding** will be identified. This procedure identifies **recombinant** vectors which can be advantageously used in the present invention.

The insertion step creates a **discrete** population of vectors, each **member of the population** containing {a nucleotide} **an oligonucleotide** insert encoding {for} a different peptide {sequence} **from a population of random amino acid sequences**, each encoded peptide sequence containing the same desired number of amino acid residues{. This}**, preferably five. The discrete** population of vectors is then transferred into a population of appropriate host cells. Concentrations of vectors and of host cells can be controlled to ensure that**, on the average,** no more than one

vector is transferred into any **individual** host cell. Cells are plated and cultured, and the translated proteins are harvested therefrom.

The population of recombinant λ bacteriophage, as described in Example IV, with random oligonucleotides inserted, will express fusion proteins containing the heterologous peptides of random amino acid sequence. In this embodiment, the heterologous peptides are located within the pIII minor coat protein. Other insertion sites may be utilized as understood by the skilled artisan for particular desired purposes. For example, Parmley and Smith (1988) Gene 73:305-318 demonstrates the expression of foreign epitopes at the N-terminal end of pIII of λ. Devlin et al. (1990) Science 249:404-406 describes a novel expression vector (Ml3LP67) derived from M13mp19; foreign epitopes were expressed near the N-terminus of the processed form of β-galactosidase. Cwirla et al. (1990) Proc. Natl. Acad. Sci. USA 87:6378-6382 reports the expression of a population of peptides expressed fused at the N-terminus of pIII of modified bacteriophage fd.

Creating the Matrix

The particular construction of the matrix created from the full range of antibodies or from the peptide sequences described above depends on its use. Either the antibodies or the peptide sequences are immobilized on a **solid support** substrate **or an immobile phase**, e.g., nitrocellulose **if a two dimensional support is desired or material which can be incorporated in a column if a three dimensional support best serves its purpose, as will be understood by the ordinary skilled artisan**. The immobilization can be accomplished by covalently linking the antibodies or peptide sequences to the substrate. Each site on the matrix is occupied by a single chemical species, i.e., a monoclonal antibody or a purified peptide. The source of each individual immobilized species is maintained as a separate culture. In general, the

antibodies**, the** peptide sequences, or the test species are labeled with an appropriate label, such as a fluorescent compound, an enzyme, or a radioactive tracer**, as known in the art**. The peptide sequence itself can serve as a sensitive biological tag where it occurs on the surface of a protein {or vector}**, virus or modified host cell**.

Where the antibodies are immobilized, the **peptide sequences or polypeptides comprising those** peptide sequences are then contacted with the antibodies under appropriate conditions and for a sufficient amount of time so that each immobilized antibody binds to the peptide sequence to which it is specific. Where the peptide sequences are immobilized, the antibodies are then contacted with the peptide sequences so that each immobilized peptide sequence is recognized and bound by an antibody specific {to} **for** that **particular** sequence. Each complex of peptide sequence and its bound antibody can be termed a binding pair. In some cases,, the antibodies or peptide sequences themselves are immobilized on the substrate; in other cases the cell cultures producing the antibodies or **the modified host cells expressing the** peptides are immobilized. Binding pairs are created in a single step, taking advantage of the natural affinity of antibodies for the peptide sequences to which they are specific. If a sample of peptides is contacted with a population of immobilized antibodies, then the peptides will self-sort and each will bind to its corresponding antibody. Similarly, if a sample of antibodies is contacted with a population of immobilized peptides, then the antibodies will self-sort and each will bind to its {corresponding} **cognate** peptide. The sorting will occur notwithstanding that there is no prior knowledge as to the functional characteristics of any of the individual antibodies or peptides.

A matrix where the antibodies are immobilized on the substrate will be designated an antibody-immobilized matrix, or AIM. Where each immobilized antibody forms a binding pair with a corresponding peptide sequence, the matrix will

be designated P-AIM. Similarly, a matrix where the peptide sequences are immobilized {on the substrate will be designated a peptide-immobilized} matrix, or PIM. Where each immobilized peptide sequence forms a binding pair with a corresponding antibody, the matrix will be designated A-PIM.

Generally, the method of the invention involves contacting a test species with an intact P-AIM or an intact A-PIM, the specific characteristics of the matrix depending on the nature of the information sought **as the skilled artisan will readily understand**. Considering the large number of different hybridomas**, recombinant vectors** and genetically {engineered clones} **modified host cells** that are involved in the {procedure} **practice** of the invention, the antibodies or peptide sequences can be immobilized very densely on the substrate. Areas of competitive binding are identified when the test species is contacted with the matrix.

{Colonies} **Recombinant vectors or modified host cells or colonies** from these areas **of competitive binding** can then be retrieved, {replated} **repeated** less densely, and the competitive binding step with the test species repeated in order to specifically identify the individual colony producing the antibody or amino acid sequence where pairing was disturbed.

Screening an Antibody or Test Species of Interest

A P-AIM is used both to identify and obtain antibody clones that are specific to a test species of interest and to identify the specific peptide sequence recognized by an antibody of interest. The test species can be, for example, a virus, a bacteriophage, a virus coat protein, a surface protein of a viral or bacterial pathogen, a protein on the surface of a malignant cell, an enzyme, or a peptide having the sequence of a selected portion of a protein of interest. The test species need not

contain peptides, but may be, e.g., a drug or carbohydrate having a {configuration} **three dimensional structure** that is closely approximated by a peptide sequence.

The test species is contacted with a P-AIM in a competitive binding assay with each of the complexed binding pairs. Each binding pair occupies a unique site on the matrix. Where these pairs have been labeled, any pairings disturbed by the presence of the test species can be identified.

A particularly sensitive labeling technique is obtained where the peptide sequences bound to the immobilized antibodies are on the surface of a protein or vector. After the P-AIM is created and the binding pairs are established, the P-AIM is thoroughly washed to remove any unbound peptide sequences. The test species is then contacted with the P-AIM. Any peptide sequences that are displaced from their corresponding antibodies by the presence of the test species can be directly titered off the P-AIM. Available techniques are sufficiently sensitive to detect the presence of as few as ten molecules of protein {or}**, recombinant** vector {organisms} **or modified host cells** in the titered supernatant.

Where the test species is labeled, its binding can be detected directly. Each clone producing an antibody that binds to a test species is identified and cultured to provide a source of the antibody. Each culture producing a peptide{,} sequence displaced by the presence of an antibody of interest is identified and cultured to provide a source of that peptide sequence.

A PIM is used both to identify the specific sequences, on a test protein or polypeptide that can be {recognize} **recognized** by antibodies and to identify the specific peptide sequences recognized by an antibody of interest. **Each clone or peptide in a PIM represents the expression or presence of at least 10^4-10^7 copies of the individual peptide sequence so that detection of labeled antibody binding or of the displacement of bound labeled antibody is readily accomplished using**

techniques known to the art. The procedure for screening on a PIM is analogous to the procedure, above, for screening on an AIM. The test protein or peptide sequence, or the test antibody, is contacted with an intact A-PIM in a competitive binding assay with each of the antibody-peptide sequence pairs. The pairings disturbed by the presence of the test protein or polypeptide or test antibody are noted, and the clones producing the amino acid sequence to which pairing was disturbed are identified and cultured. By this method, not only is it possible to determine the amino acid sequence recognized by the antibody, but it is now possible as well to identify {the} **a** nucleic acid sequence encoding this amino acid sequence, as the **oligonucleotide** insert in the vector contained in the clone that produces the recognized amino acid sequence.

{Example I} **EXAMPLE 1**

To illustrate certain aspects of the present invention, a method for determining the antibody recognition sites on insulin {will now be} **is** described.

Production of {hybridoma cell lines} **Hybridoma Cell Lines**

Several C57Bl/10 mice are **each** immunized intraperitoneally with 100 micrograms of human insulin precipitated in alum, mixed with {2xl09} **2×10^9** killed *{Bordatella}* ***Bordetella*** pertussis organisms as {adjuvent} **adjuvant**. A second injection of 100-200 micrograms of insulin in saline is given a month later.

Three days after the second injection, {the mice are killed by neck dislocation,} the spleens are removed aseptically and transferred into a **sterile** bacteriological-type plastic petri dish containing 10 ml of GKN solution. GKN solution contains, per 1 liter of distilled water: 8 g {NaCl} **NaC1**, 0.4 g KC1, 1.77 g {Na2HPO4·2H2O,} **$Na_2HP0_4 \cdot 2H_20$,** 0.69 g {NaH2PO4·H2O} **$NaH_2PO_4 \cdot H_2O$**, 2 g

glucose, and {0.01} **0. 0l** g phenol red. The cells are teased from the capsule with a spatula. Clumps of cells are further dispersed by pipetting up and down with a 10 ml plastic pipette. The suspension is transferred to a 15 ml polypropylene tube where clumps are allowed to settle for 2 to 3 minutes. The cell suspension is decanted into another tube and centrifuged **at 170 x G** for 15 minutes {at 170 G} at room temperature. The cells are washed again in GKN and {finally} **then** resuspended in 1-2 ml GKN. A 20 microliter aliquot of the cell suspension, stained with 1 ml of trypan blue solution, is counted to determine the yield of spleen cells.

10^8 washed spleen cells and 5 x 10^7 8-azaguanine resistant myeloma cells (e.g., cell line X63Ag8.6.5.3; FO; or Sp2/0-{Agl4)} **Ag14)** are combined in a 50 ml conical tube (Falcon {19-}2070). The tube is filled with GKN and {spun} **centrifuged** at 170-200 G at room temperature. The supernatant is {then} withdrawn, and 0.5 ml of a 50% solution of polyethylene glycol in GKN is added dropwise to the pellet. This addition is accomplished over a one minute period at room temperature as the pellet is broken up by agitation. After 90 seconds 5-10 ml of GKN are added slowly over a period of 5 minutes. The cell suspension is then left for 10 minutes, after which large clumps of cells are dispersed by gentle pipetting with a 10 ml pipette. The cell suspension is then diluted into 500 ml of {Dulbecco's} **Dulbecco's** modified Eagles medium containing 10% fetal calf serum and HAT. 1 ml aliquots are distributed into 480 wells of Costar-Trays (Costar Tissue Culture Cluster 24, Cat. No. 3524, Costar, 205 Broadway, Cambridge, MA) **each well** already containing 1 ml HAT medium and {105} **10^5** peritoneal cells or 10^6 spleen cells. The trays are kept in a fully humidified incubator at 37°C in an atmosphere of 5% CO_2 in air. After 3 days and twice a week {2} thereafter, 1 ml medium is removed from each well and replaced with **fresh** HAT medium. After 7-10 days the wells are inspected for hybrids and the HAT medium is replaced with HT medium. Cell populations of

interest are expanded by transfer into cell culture bottles for freezing, cloning, and product analysis. 10^6 peritoneal cells are added at this time to each culture bottle.

Hybridomas produced by the methods outlined above are propagated and cloned, using standard techniques. The monoclonal antibody produced by each hybridoma line is purified from the culture supernatant and concentrated by affinity chromatography on a protein A-sepharose column.

Production of {gene library} **Gene Library**

cDNA is synthesized from a heterogeneous population of mRNA**, prepared from bovine pancreas**. The cDNA is randomly sheared and the 15 nucleotide fragments are retrieved by electrophoresis. These fragments are inserted, in phase, into the structural gene encoding beta-galactosidase of {lambda-gt 11} **λgtll**, according to the {method} **methods** disclosed in Pieczenik, U.S. Patent 4,359,535{. Each of the resulting clones produces the normal lambda-gt 11} **and 4,528,266. Cells infected with each of the resulting recombinant bacteriophages produce the normal λgt11 proteins plus a hybrid beta-galactosidase** protein containing a foreign sequence of 5 amino acid residues encoded by the 15 nucleotide fragment inserted into the beta-galactosidase {gene.}**(lacz) gene. From 1 microgram of double-stranded oligomeric DNA, 15 base pairs in length, about 6×10^{10} individual sequences can be cloned if cloning is 100% efficient.**

{Screening and precise identification} **Screening and Precise Identification {of the antibody binding} of the Antibody Binding** Sites

The library is plated at a density of 25,000 plaques per {150-mm2} **150CM²** plate and immunologically screened, using a pool of those monoclonal antibodies reactive with human insulin and unreactive with unmodified {lambda-gt 11} **λgt11**

phage. The immunological screening is carried out essentially according to the method described by Young{, R.A.} et al., Science (1983) 222{,}:778, which is hereby incorporated by reference.

The {lambda-gt 11} **recombinant λgt11** clones identified by the screening procedure are introduced as lysogens into E. coli strain RY 1089 (ATCC 37,196). Lysogens are grown at 32°C in media containing 50 micrograms of ampicillin per milliliter {until absorbance} **to an optical density** at 550mm {is} **of** 0.4 to 0.8. The {phages} **recombinant phage** are induced at 44°C by shaking gently for 20 minutes and then {isopropyl-thiogalactoside} **isopropylthiogalactoside** (IPTG) is added to a final concentration of 2mM, and the culture is shaken an additional hour at 37°C in order to {enhance} **induce** expression of **hybrid** beta-galactosidase and possible fusion proteins.

Lysates are then {subjected to electrophoresis on a sodium dodecyl sulfate-polyacrylamide gel }**analyzed by sodium dodecyl sulfatepolyacrylamide gel electrophoresis** (SDS-PAGE) and electroblotted {into} **onto** nitrocellulose. Pelleted cells from 0.1 ml of each lysogen culture are {dissolved} **suspended** in 20 microliters of SDS gel sample buffer (3% SDS, 10% glycerol, {10 mM} **10mM** dithiothreitol, 62 {mM tris} **MM Tris**-HC1, pH 6.8) **and proteins are solubilized** at 95°C for 5 minutes {for electrophoresis. }**before electrophoresis. Proteins are separated by SDS-PAGE according to the method of Laemmli (1970) Nature 277:680 with a 4.5% stacking gel and an 8-12% gradient gel.** Western blot analysis is performed according to a modification of the method of Towbin H. et al. (1979) Proc. Natl. Acad. Sci. U.S.A. 79 {4350. Proteins are separated by SDS-PAGE according to the method of Laemmli (1970) Nature 277 680 with a 4.5% stacking gel and an 8-12% gradient gel. The}**:4350. Each** filter is reacted for 90 minutes with a single one of the monoclonal antibodies selected above diluted to a concentration of 1:20,000 with

PBS containing 0.05% Tween-20 and 20% FCS. Filter-bound antibody is incubated with [^{125}I]-labeled sheep antiserum prepared against whole mouse antibody (diluted to 2 x 10^5 cpm/ml with PBS containing 0.05% Tween-20 and 20% FCS) and then detected by autoradiography. The lysogen that is reactive with the specific antibody used contains the engineered {lambda-gt 11} **λgt11** clone whose beta-galactosidase enzyme is fused to a 5 amino acid sequence that corresponds to the 5 amino acid sequence of insulin recognized by {the} **that** antibody. The electrophoresis and electroblotting steps are repeated for each of the monoclonal antibodies selected above, and the specific sequences on the insulin molecule recognized by each of these antibodies {is identified. }**are identified by determining the DNA sequences of the oligonucleotide inserts and deducing the respective encoded amino acid sequences.**

{Example} **EXAMPLE II**

The method of Example 1 is modified to eliminate the step of {innoculating} **inoculating** the mice with human insulin. An identical harvesting procedure is used to obtain spleen cells from mice that have not been antigenically stimulated. The spleen cells are hybridized with myeloma cells as described in Example 1, and the resulting hybridomas are propagated and cloned. Notwithstanding the elimination of the antigenic stimulation step, screening identifies clones that produce {:} antibodies reactive with human insulin.

{Example} **EXAMPLE III**

To further illustrate the invention, a method for creating and screening a cDNA expression library will now be described. In this example, the cDNA library is prepared from chicken smooth muscle mRNA.

Production of Gene Library

Total smooth muscle RNA is prepared from 11-day embryonic chicken stomachs and gizzards according to the {!}method of Chirgwin, J.M. et al., (1979) Biochemistry 18:5294 and Feramisco, J.R. et al., (1982) J. Biol. Chem. 257:11024.

Poly {(A)}**(A)** + RNA is isolated by two cycles of adsorption to and elution from oligo(dT)-cellulose according to the method of Aviv, H. et al., (1972) Proc. Natl. Acad. Sci. USA 69:1408. Starting with about 25 micrograms of poly {(A) +RNA}**(A) + RNA**, first and second strand cDNA is synthesized using avian myeloblastosis virus reverse transcriptase. The double linker method of Kartz and {Nicodemus} **Micodemus**, (1981) Gene 13:145 can be employed. The double stranded cDNA, with intact hairpin loops at the ends corresponding to the 5' ends of the poly(A)+ mRNA, are filled in with the Klenow fragment of E. coli DNA polymerase I (available**, e.g.,** from Boehringer Mannheim or New England BioLabs). The filled in cDNA is then ligated to [^{32}P{-}]-labeled {Sal I} **SalI** octanucleotide linkers (available from Collaborative Research, Waltham MA). The cDNA with {Sal I} **SA1I** linkers attached to the end corresponding to the 3' end of the poly(A)+ mRNA is then treated with nuclease Sl to destroy the hairpin loop and again is filled in with the Klenow fragment of E. coli DNA polymerase I. EcoRI octanucleotide linkers {(also available from}(Collaborative Research) are ligated to the cDNA. The {DNA} **cDNA** is digested to completion with both EcoRI and {Sal I} **Sa1I**. A Sepharose 4B column equilibrated with 10mM Tris-{HCl} **HC1** (pH 7.6) containing 1 mM EDTA and 300 mM NaCl

is used to isolate and purify those **CDNA fragments containing oligonucleotide sequences, 15 nucleotides in length, which are then flanked by the octanucleotide linkers.** {cDNA fragments 15 nucleotides long flanked by the two octanucleotide linkers.}

The plasmid vector pUC8, described in {Vieria} **Vieira** et al. **(1982)** Gene 19{(1982)}**:**259, is digested to completion with EcoRI and {Sal I} **SalI** and extracted twice with a 1:1 {by volume}**(v/v)** mixture of phenol and chloroform. The 2.9 kilobase fragment is separated from the {16 nucleotide long} **oligonucleotide** fragment on a Sepharose {413} **4B** column, equilibrated as set forth above. Fractions containing the large fragment are pooled and precipitated with ethanol. cDNA is ligated to the vector at a weight ratio of vector to cDNA of 1000:1. Approximately 1 nanogram of cDNA is ligated to 1 microgram of the plasmid vector.

Conventional techniques are employed to transform E. coli strain DH-1 with the engineered pUC8 vector. The {bacteria} **transformed bacterial cells** are plated onto 82 mm nitrocellulose filters (Millipore Triton-free HATF) overlaid on {amplicillin} **ampicillin** plates to give about 1,000 colonies per filter. Colonies are replica plated onto nitrocellulose sheets (available from Schleicher & {Schnell)} **Schuell)** and the replicas are regrown both on selective plates for antibody and hybridization screening and on glycerol plates for long-term storage at -70°C.

Antibody Production and Immunological Screening

Each plate is immunologically screened to identify colonies where the plasmid contains a 15 {nucleotide cDNA insert }**base pair oligonucleotide insert encoding a peptide sequence** corresponding to a portion of the chicken tropomyosin gene. Monoclonal antibodies for use in the screening are developed as follows.

Spleen cells are harvested from donor mice that have been antigenically stimulated with chicken tropomyosin. Alternatively, spleen cells {are} **can be** harvested from mice that have not been antigenically stimulated. The spleen cells are fused to myeloma. cells to produce hybridoma strains. The monoclonal antibody

produced by each hybridoma line is purified from the culture supernatant and concentrated by affinity chromatography on a protein A sepharose column.

Antibodies are screened for reactivity with chicken tropomyosin and with the parental bacterial strain, DH-1, {which does not contain a plasmid} **preferably containing unmodified pUC8**. Those antibodies reactive with the tropomyosin and unreactive with DH-1 **(pUC8)** are selected for use in screening the transformed bacterial colonies.

To prepare the bacterial colonies for screening, {they} **cells** are lysed by suspending the nitrocellulose filters for fifteen minutes in an atmosphere saturated with $CHC1_3$ vapor. Each filter is then placed in an individual Petri dish in 10 ml of 50 mM Tris-HC1{, pH 7.5/150} **(pH 7.5) 150** mM {NaCl/5} **NaCl, 5** mM {MgCl2} **$MgC1_2$** containing 3% (wt/vol) bovine serum albumin, 1 microgram of DNase, and 40 micrograms of lysozyme per milliliter. Each filter is agitated gently overnight at room temperature, and then rinsed in saline (50 mM Tris-HC1, {pH 7.5/150}**(pH 7.5) 150** mM NaCl). Each filter is incubated with a dilute saline solution of a monoclonal antibody selected from those antibodies exhibiting reactivity with tropomyosin but not with DH-1 **(pUC8)**. The filters then are washed five times with saline at room temperature, {from} **for** one half to one hour per wash. The filters then are incubated with 5 x 10^6 cpm of **[^{125}I]**-labeled goat anti-mouse IgG {having} **at** a specific activity of about 10^7 cpm/microgram {and} diluted in 10 ml of saline containing 3% bovine serum albumin. The goat anti-mouse IgG can be an affinity purified fraction. The labeling is accomplished according to the chloramine-T procedure of Burridge, K. (1978) Methods Enzymol. 50:57. After one hour of incubation the filters are washed again in saline, with five or six changes, at room temperature, dried, and autoradiographed 24-72 hours**, preferably** using Dupont Cronex Lightning Plus x-ray enhancing screens. In the immunological screenings, a

filter is advantageously included upon which defined amounts of various purified proteins are spotted. This serves as a further control for the specificity of the immunological detection of the antigens. Quantities of less than 1 nanogram of purified protein can be detected in these assays.

This procedure permits the identification and characterization of the specific five {peptide} **amino acid epitopic** sequence of the tropomyosin protein that is identified by a particular monoclonal antibody. As this immunological screening process is repeated with different monoclonal antibodies, several distinct antigenic sites on the tropomyosin protein are identified. The 15 nucleotide sequence of cDNA that encodes {for} each antigenic site is preserved in the cDNA-**derived** library, and a source of antibody that recognizes each site is preserved in the separate hybridoma lines.

Use

The invention is useful to produce antibodies that recognize and bind to particular test species, and to determine either (1) the specific peptide sequence on a protein, enzyme, or peptide that an antibody recognizes or (2) an amino acid sequence with a configuration very close to the structure of a non-peptide **or a discontinuous epitopic** test species recognized by an antibody. The invention, is also useful to determine the nucleotide sequence **or sequences according to the codon degeneracy,** encoding the amino acid sequence that is recognized by an antibody.

To identify a peptide sequence that closely approximates an antibody binding site on a test species, either an A-PIM or a P-AIM can be used. If an A-PIM is used, then the test species is first contacted with the intact A-PIM. Any antibodies bound to immobilized peptide sequences that have an affinity for the test species will be "competed off" the matrix to bind to the test species. The peptide sequence

immobilized at a site where antibodies are "competed off" has a conformational similarity to the site on the test species where the antibodies are now bound. If a P-AIM is used, then the test species is first contacted with the intact P-AIM. The test species displaces any peptide sequences that have a sufficient conformational similarity to an antibody recognition site on the test species that an antibody capable of binding to the peptide sequence is also capable of binding to the test species. Displaced peptide sequences can then be titered off the matrix and identified. It is not necessary that the test species be {proteinaceous} **protainaceous** or derived from peptides. It can be, for example, a carbohydrate or a non-peptide drug. It can be expected that the recognition site of a non-peptide substance {is} **can be** closely approximated by the conformation of a peptide sequence **or that a linear amino acid sequence can mimic a discontinuous epitope**. A test species can disturb the binding at more than a single site on a matrix; this could occur because there is more than one distinct antibody recognition site on the test species or because two {or} **ore** more distinct peptide sequences are each similar in conformation to a {recognition site} **single epitope** on the test species. It is not necessary that the test species be immunogenic, i.e., induce the production of antibodies in vivo if {innoculated} **inoculated** into a mammal; the antibody binding sites of a test species can be characterized {notwithstanding} **even when** that {the} test species is not immunogenic.

Where the test species is a disease producing agent, such as a virus or {bacteria} **a bacterium**, then the peptide sequences that are similar in conformation to the antibody recognition sites of the disease producing agent can be employed to develop a vaccine. A synthetic antigen incorporating the identified peptide sequence or sequences, when injected into a patient's bloodstream, {induces} **can induce** the production of antibodies against the disease producing agent.

Where the test species is the recombinant gene product of a gene expression library, one {is able to} **can** determine precisely what regions of the gene product make up antibody recognition sites. The identified peptide sequences{,} correspond to sequences contained in the gene product that are recognized by antibodies.

Where the test species is a gene product, such as, for example, a protein, **an** enzyme, or **a** peptide, then the invention also provides a means for locating in a genome the gene encoding {for the} **that** gene product. After the peptide sequences identified from screening the gene product through the matrix are identified, the recombinant cell lines that produced those peptide sequences are identified and the {recombinant nucleotide} **oligonucleotide** sequences encoding those peptide sequences are {recovered} **determined**. The nucleotide sequences can then be used as {a} DNA {probe} **probes** to locate on the genome the gene encoding for the gene product. Since each nucleotide sequence is fairly short, i.e., from **about** 5 to **about** 12 triplets in length, it can be expected that any one sequence, or a closely similar sequence, would be repeated more than once in the genome. Therefore, several distinct nucleotide sequences, each encoding a distinct peptide sequence, are advantageously employed in {a} DNA {probe} **probes**. A region on a chromosome where several nucleotide sequences hybridize in close proximity identifies **the DNA fragment containing** the gene encoding for the gene product.

To determine the peptide sequence recognized by a particular antibody of interest, either a PIM or a P-AIM can be used. If a PIM is used, it is not necessary that each immobilized peptide sequence be bound to a corresponding antibody. The antibody of interest can be contacted directly with a matrix of immobilized peptide sequences. Any immobilized sequences that are bound by the antibody of interest can then be directly identified. If a P-AIM is used, then the antibody of interest is first contacted with the intact P-AIM. Any peptide sequences bound to immobilized

antibodies that can be recognized by the antibody of interest will be "competed off" the matrix to bind with the antibody of interest. Peptide sequences that have been "competed off" the matrix by the presence of the antibody of interest can then be titered off the matrix and identified.

Where it is desired to determine the nucleotide sequence encoding the peptide sequence recognized by an antibody of interest, the **modif ied host cell or** recombinant {cell line} **protein or virus** that produces the peptide sequence recognized by the antibody can be identified and the nucleotide sequence encoding the peptide sequence can be recovered and {sequenced} **the sequence can be determined**.

Where the antibody of interest is an antibody produced by a patient suffering from an autoimmune disease and the antibody attacks the patient's own cells **or protein**, impairing the functioning of those cells **or protein**, then the peptide sequence recognized by the antibody can provide a basis for treating the patient. The peptide sequence recognized by the antibody can be administered to the patient in an **amount** effective {amount to} **for** competitively {inhibit} **inhibiting** the antibody from attacking the patient's own cells **or protein** in vivo. The patient's condition will be improved since fewer antibodies will be available to attack the living cells{, and the} **or functional protein. The** administration of peptides will not induce further antibody production since the peptides are too short to induce an immunogenic response.

To identify an antibody that reacts with a test species, an AIM is used. It is not necessary that each immobilized antibody be bound to a corresponding peptide sequence. The test species can be contacted directly with a matrix of immobilized antibodies. Any immobilized antibodies bound to the test species can be directly identified, and the clones producing those antibodies can be cultured to provide a

source of the antibodies. It is not necessary that the test species be proteinaceous or derived from peptides. It can be, for example, a carbohydrate or a non-peptide drug. It is not necessary that the test species be immunogenic. It is possible to obtain antibodies that recognize a test species {notwithstanding that} **even though** the test species**, itself,** does not induce antibody production in vivo.

The antibodies that recognize the test species can be used in an immunoassay to test for the presence of the test species in a biological sample.

Where the test species is associated with a disease, then an antibody **(or antibodies)** that recognizes the test species can be used in a diagnostic test kit to determine the condition of a patient. The antibody**(ies)** is contacted with an appropriate sample from the patient to test for the presence of the test species, which is associated with a particular disease. The antibody**(ies)** can be incorporated into a diagnostic test kit that recognizes {an epitope} **one or more epitopes** on a disease-associated substance.

Where the test species is a population of malignant cells from a patient, e.g., cancer cells, then an antibody that **specifically** recognizes the malignant cells while not recognizing healthy cells from the patient can be used to target drugs to the malignant cells. A sample of malignant cells is contacted with an AIM and antibodies that bind to the malignant cells are identified. A sample of healthy cells from the patient is contacted with a replica of the matrix, and antibodies that bind to the malignant cells, but not to the healthy cells, are selected. A hybridoma line producing selected antibodies is cultured to provide a source of the selected antibodies. A drug, {e.g.,} **or other** cytotoxic agent, is then linked to the selected antibodies, and {an} **a therapeutically** effective amount of the drug-linked antibodies is administered to the patient.

EXAMPLE IV

This example demonstrates the incorporation of a discrete population of oligonucleotides encoding a population of peptides, each peptide comprising five amino acids in random order, into the λ gene encoding the minor coat protein pIII (gene III). Thus, a discrete population of recombinant vectors was produced.

The universe of peptides of random sequence, each five amino acids in length, is 5^{20}, or 3.2×10^6.

One way to generate each pentapeptide sequence is to take advantage of the fact that a population of random nucleotide sequences, each 15 nucleotides in length, can encode the population of random peptide sequences each five amino acids in length.

Because the genetic code is degenerate, i.e., there are 61 codons coding for 20 amino acids; each amino acid, on the average, has 61/20 or 3.05 synonymous codons. In terms of the nucleotide universe, there are 61 to the power 5 possible nucleotide sequences coding for the 3.2 million pentameric epitopes. Therefore, there are 844,596,301 possible nucleotide sequences coding for 3,200,000 possible pentapeptide sequences. This means that there are 263.94 synonymous codings for each pentapeptide sequence. This high degree of synonymous degeneracy allows us one way of evaluating whether one has generated the universe of possible pentameric epitopes. Generating 3-5 synonymous representations of the coding for the pentapeptide universe statistically suggests an almost complete representation of each member of the pentameric universe. That is, if the nucleotide distribution generated is equimolar and random, one would expect that if one randomly generated 3-5 synonymous codings for any particular pentameric peptide sequence, one would have had a statistically good chance of

having generated any other pentameric peptide sequence in the population of 3.2 million possible pentamers.

A discrete population comprising a random distribution of nucleotide sequences (15 mers) and thereby at least one copy of each of the sequences encoding all possible pentapeptides was chemically synthesized as oligonucleotides of the formula GATCCTTN$_{15}$ AA SEQ ID NO: 1 where N is G, A, T or C. The 15 base random sequences are the coding sequences for the peptide epitope universe 4^{16} or 4, 294, 967, 296 different molecules were synthesized at an average of 243 codings per pentapeptide sequence, this represents a population with about five-fold redundancy. About 1 microgram of DNA was recovered and 10^8-10^9 recombinant phage were produced. The TT and AA bases at the 5' and 3' ends, respectively, will allow the sequence to base pair with itself in phase on both strands if GAT is in the sense phase. In addition, the oligonucleotide, after hybridizing to a complementary oligonucleotide, can be ligated in a BamHI site without regenerating a BamHI site so that a BamHl selection against parental molecules lacking inserts can be performed.

One test of the randomness of the chemical synthesis is that half of the approximately 4.2 x 10^9 oligonucleotides should be able to form duplexes with the other half. The oligonucleotides were purified on a Sep Pak™ (Millipore, Waters Chromatography, Milford, MA) column, lyophilized and resuspended in ligation buffer, heated at 100°C 5 min and brought to room temperature slowly and incubated overnight. The duplexed oligonucleotides were then ligated into λ RF DNA which had been previously digested with BamHI and purified after a-garose gel electrophoresis. The ligation mixture was transfected after BamHI digestion into freshly prepared competent E. coli TGI cells and plated essentially

as described (Smith (1985) Science 228:1315-1317). E. coli TGI is a RecA derivative of E. coli JM101.

Representative plaques were picked and screening-using only one sequencing track to identify bacteriophage with inserts. About one-third of the plaques screened were derived from bacteriophage-carrying inserts. These recombinant bacteriophage were plaque-purified and the inserts were sequenced essentially as described in de la Cruz et al. (1988) J. Biol. Chem. 263:4318-4322. Table 1 shows the oligonucleotide sequences of fourteen randomly chosen inserts. The accompanying statistical analysis shows that the observed base distribution is not significantly different from the expected random (equimolar) distribution of bases. Thus, it was confirmed that random oligonucleotides could be synthesized, a particular oligonucleotide could find its complement (or one sufficiently similar to allow duplexing) and that a sequence inserted in an epitopic λ vector was stable. Furthermore, the recombinant bacteriophages were viable.

Several of the amino acid sequences encoded by the fourteen random oligonucleotides of Table 2 are also found in databases of protein sequences (Genbank, Atlas of Protein Structure and Sequence) at a frequency expected for a codon distribution determined by random nucleotide sequences. The fourteen translated sequences and the proteins containing identical amino acid sequences are given in Table 2.

TABLE 1

Base Composition Analysis of Randomly Synthesized Coding for Epitopes

1)	CTT ACCGAGCGGACTGGT AAA	
2)	CTT ATGCAAGACTCGATA CAK	
3)	CTT GCGGGGTCAGAGGGC GAA	
4)	CTT CAGATATTTCCGAAG CAA	
5)	CTT AACATCCTCCAACGG CAA	
6)	CTT CCATCGCTGAAACTC AAA	
7)	CTT ACACCGAGGGCGCTC CAN	
8)	CTT CTAGAATTCGTGGGC AAA	
9)	CTT AGCGTGCTCGACAGG CAA	
10)	CTT CAAGACAAAGTACAT CAA	
11)	CTT GAAGTATATCAAGCA CAA	
12)	CTT GTTTTCCTTACTCCC GAA	
13)	CTT CTATACATAACCAAC AAA	
14)	CTT GACGCGGATATAGGA AAA	
T)	0 4 1 3 5 0 4 7 4 1 3 2 0 3 2 0	= 39
C)	5 4 4 2 3 4 5 1 4 3 6 2 5 2 6 6	= 50
G)	4 1 3 7 1 6 2 2 3 5 0 5 5 6 3 3	= 53
A)	5 5 6 2 5 4 3 4 3 5 5 5 4 3 3 5	= 62
Totals	14 x 15 = 210	210

Coding Strand Composition Phage Strand(+) Composition

T=39/210=18.6%	= A
C=56/210=26.7%	= G
G=53/210=25.2%	= C
A=62/210=29.5%	= T

T	25-18.6 = 6.4 X	6.4 = 40.96/25	=1.64
C	25-26.7 =-1.7 X	1.7 = 2.09/25	= .18
G	25-25.2 =-0.2 X	-0.2 = .04/25	= .04

A $\frac{25\text{-}29.5 = -4.5 \text{ X } -4.5 = 20.25/25}{X^2}$ $\frac{= .81}{= 2.57 \text{ ; } 3D.F.}$

A distribution with a Chi-square of 2.57 and 3 degrees of freedom can be gotten randomly 50% of the time. Therefore, our observed distribution does not differ significantly from our expected (and synthesized) global base composition.

TABLE 2

1) TCTTTACCAGTCCGCTCGGTAAGATCCTCA
TGAGGATCTTACCGAGCGGACTGGTAAAGA
THRGLUARGTHRGLYLYS
T E R T G K

Phaseolin-Kidney bean

2) TCTTGTATCGAGTCTTGCATAAGATCCTCA
TGAGGATCTTATGCAAGACTCGATACAAGA
M Q D S I Q

3) TCTTCGCCCTCTGACCCCGCAAGATCCTCA
TGAGGATCTTGCGGGGTCAGAGGGCGAAGA,
A G S E G E

4) TCTTGCTTCGGAAATATCTGAAGATCCTCA
TGAGGATCTTCAGATATTTCCGAAGCAAGA
Q I F P K Q

5) TCTTGCCGTTGGAGGATGTTAAGATCCTCA
TGAGGATCTTAACATCCTCCAACGGCAAGA
N I L Q R Q

Fibrinogen gamma B chain precursor
Fibrinogen gamma A chain precursor

6) TCTTTGAGTTTCAGCGATGGAAGATCCTCA
TGAGGATCTTCCATCGCTGAAACTCAAAGA
P S L K L K

P3 protein-Bluetongue virus
H-2 class 1-related secreted histocompatability

7) TCTTGGAGCGCCCTCGGTGTAAGATCCTC
TGAGGATCTTACACCGAGGGCGCTCCAAGA
T P R A L Q

RNA-directed RNA polymerose

8) TCTTTGCCCACGAATTCTAGAAGATCCTCA
TGAGGATCTTCTAGAATTCGTGGGCAAAGA
L E F V G K

9) TCTTGCCTGTCGAGCACGCTAAGATCCTCA
TGAGGATCTTAGCGTGCTCGACAGGCAAGA
SERVALLEUASPARGGLN
S V L D R Q

Coat protein-Cauliflower mosaic virus
Anthranilate synthase

10) TCTTGATGTACTTTGTCTTGAAGATCCTCA
TGAGGATCTTCAAGACAAAGTACATCAAGA
Q D K V H Q

Beta casein-bovine

11) TCTTCTGCTTGATATACTTCAAGATCCTCA
TGAGGATCTTGAAGTATATCAAGCAGAAGA.
E V Y Q A E

Nucleoapsid protein N-Punta Toro phlebovirus
Tyrosine amino transferase-rat

12) TCTTCGGGAGTAAGGAAAACAAGATCCTCA
TGAGGATCTTGTTTTCCTTACTCCCGAAGA
V F L T P E

Pol polyprotein-Bovine leukemia virus

13) TCTTTGTTGGTTATGTATAGAAGATCCTCA
TGAGGATCTTCTATACATAACCAACAAAGA
L Y I T N K

14) TCTTTTCCTATATCCGCGTCAAGATCCTCA
TGAGGATCTTGACGCGGATATAGGAAAGA

D A D I G K

EXAMPLE V

Rabbit polyclonal antibodies specific for the N-terminus of endoplasmin were prepared as described herein.

A peptide containing the N-terminal fifteen amino acid residues of endoplasmin, with an added C-terminal tyrosine residue, is synthesized as described (Cameron et al. (1987) J. Chem. Soc. Chem. Commun. 0(4):270-272). The sequence synthesized is

Asp-Asp-Glu-Val-Asp-Val-Asp-Gly-Thr-Val-Glu-Glu-Asp-Leu-Gly-Tyr.

The synthetic peptide was coupled, in separate reactions, to keyhole limpet hemocyanin (KLH) and bovine serum albumin (BSA) in a ratio of 5mg peptide to 30mg carrier protein (KLH or BSA), using bis-diazotized o-tolidine. The peptide was suspended at a concentration of 5mg/ml in 0.16M sodium borate, 0.9% NaCl (pH 9.0). The protein was suspended at 30mg/ml in the same buffer. The o-tolidine was diazotized by dissolving 0.23g o-tolidine HC1 in 45ml 0.2M HC1, and adding 0.75g sodium nitrate in 5ml water. The mixture was stirred at 4°C for 60 min, aliquots were then stored at -20°C.

To conjugate peptide to carrier protein, 5mg peptide, 15mg protein and 0.6ml bis-diazotized o-tolidine were mixed, the volume was adjusted to 4ml and the pH was adjusted to 7.4. The reaction was carried out in the dark at 4°C for 2 hr. Excess reagents were removed by dialysis at the 4°C (against 5 1 water for 4h; against 5 1 PBS overnight). Peptide conjugates were stored in 50% glycerol in PBS (vol/vol) at -20°C.

Rabbit antisera were produced by injecting 5mg peptide-protein conjugate in 2ml 50% Freunds adjuvant every 14 days until an antibody response was detected using standard techniques.

Antibody specific for the peptide-protein conjugate was affinity purified from immune sera using a KLH-peptide strip prepared as described in Smith et al. (1984) J. Cell Biol. 99:20-28.

Defining the Endoplasmin Epitope

Peptides were chemically synthesized, each of which was a contiguous five amino acid sequence from the N-terminal amino acid sequence of endoplasmin. These peptides were immobilized to a solid support in individual spots. Polyclonal antibodies (as described above) were allowed to bind to the immobilized peptides. Detection of the bound antibody revealed that only the peptide comprising amino acids 2-6 of endoplasmin bound antibody molecules.

Recombinant phage with the chemically synthesized 15 bp oligonucleotide encoding the known epitope (amino acids 2-6) of endoplasmin with BamHI-compatible ends are prepared by inserting the coding sequence into BamHI-cut f1 RF.

Recombinant phage are propagated in liquid culture and partially purified from cell-free supernatants by three cycles of polyethylene glycol-salt precipitation and resuspension. The final supernatant is spun at high speed (about 100,000 x G) to pellet the phage. The gelatinous phage pellet (containing about 10^{11} - 10^{12} phage) is resuspended in about 50 microliters 0.2% Ponceau S in 6% acetic acid. Glycerol and tracking dye are added to make the sample sufficiently dense for gel loading. The resuspended phage mixture is then loaded onto an SAS-polyacrylamide gel and electrophoresed (Laemmli et al. (1970) supra).

After electrophoresis, the proteins in the SDS-polyacrylamide gel are transferred to nitrocellulose using standard techniques. The nitrocellulose blot is then soaked briefly in 0.2% Ponceau S in 6% acetic acid to visualize protein

bands. The pIII band is relatively sharply resolved. Then the stained blot is rinsed in water or PBS to. remove the stain. Then Western blotting is carried out essentially as described in McCafferty et al. (1990) Nature 348:552-554 with the use of Cadbury's brand of skim milk powder.

The inventors note that treatment of the phage in 6% acetic acid prior to electrophoresis is crucial for obtaining successful electropherograms and Western blots. With the acid pretreatment, recombinant phage carrying only one copy of an oligopeptide epitope can be successfully detected by Western blotting.

For topological mapping, an oligonucleotide comprising a sequence encoding amino acids 2-6 of endoplasmin as a tandem repeat of two copies, is chemically synthesized, e.g., using automated DNA synthesis (Model 380B, Applied Biosystems, Inc., Foster City, California). After synthesis and purification, the two strands of the oligonucleotide are allowed to self anneal, appropriate linkers are added, and then inserted into randomized linear λ RP molecules as previously described (U.S. Patent Nos. 4,528,266 and 4,359,535 which are incorporated by reference herein).

The recombinant λ DNA molecules are transfected into competent E. coli cells, and plated. Plaques which result from recombinant phage are identified using conventional hybridization techniques.

Phages are also screened with the endoplasmin-specific antibody described above and labelled second antibody. The immunological screening was carried out essentially as described in McCafferty et al. (1990) supra, except that the nitrocellulose containing the "lifted" plaques was first treated with 0.2% Ponceau S in 6% acetic acid for 3-4 minutes, followed by rinsing in water until destained. As before, Cadbury brand of skim milk powder is used. Isogenic E. coli transfected with unmodified λ were used as a control in the immunological

screen. Recombinant λ expressing the endoplasmic epitope comprising the pentapeptide sequence are identified by the screen.

For best results when using the BamHI site within the pIII gene for epitope analysis, one should use either a tandem repeat of at least two copies of each pentapeptide sequence encoded, or a single copy of a random pentapeptide target sequence should be flanked with a short oligopeptide sequence, e.g., about three amino acids on either side. This extra peptide sequence associated with the target sequence improves the accessibility of the epitope to antibody for binding. Similarly, the Ponceau S-acetic acid pretreatment of proteins to be blotted allows one to detect epitopes whose coding oligonucleotides are incorporated at the BamHI site within the gene encoding pIII of λ. In topological mapping or in immunological screening of plaque lifts on nitrocellulose, the acid treatment is also key to successful results.

Other Embodiments

Other embodiments are also within the {following} scope of the appended claims.

For example, it is not necessary that the matrix be constructed by immobilizing the antibodies or the amino acid sequences on a substrate. Each {clone} hybridoma producing an antibody can be cultured separately, and each {clone producing a peptide sequence} recombinant can be cultured separately. Each antibody is tested individually with each peptide sequence. Correspondence between individual antibodies and the peptide sequences{,} recognized by them can be recorded. A test species can then be tested against each of the individual antibody producing cultures. Any antibodies that bind to the test species can be identified, and the specific peptide sequence recognized by the antibody can be determined by the

corresponding peptide sequence-producing culture. Similarly, a test antibody can be tested against each of the individual peptide sequence producing cultures. The specific peptide sequence or sequences recognized by the test antibody can be determined directly by characterizing the unique peptide sequence produced by any cultures that show a positive binding response with the test antibody. This general method can readily be applied to any of the specific uses of a matrix set forth above.

In a further alternative embodiment of the invention, a submatrix can be created containing those antibody-peptide sequence binding pairs that are reactive with a test species of interest. The test species can be a peptide, enzyme, protein, a non-peptide drug, or other {non-peptide} **nonpeptide** bioactive substance. The test species is screened on a matrix containing a full range of antibodies and peptide sequences. Those antibody-peptide sequence binding pairs reactive with the test species are selected to form a submatrix. The submatrix is useful in further investigation of the immunological and conformational properties of the test species.

WHAT IS CLAIMED IS:

1. A population of oligonucleotides, wherein:

The skilled artisan will understand that any of the aforementioned vectors may be substituted or that other vectors known in the art may be used, providing sequences can be inserted in frame and that expressed random epitopes are expressed in such a way that they are accessible for antibody screening.

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: Pieczenik, George

(ii) TITLE OF INVENTION: METHOD AND MEANS FOR SORTING AND IDENTIFYING BIOLOGICAL INFORMATION

(iii) NUMBER OF SEQUENCES: 44

(iv) CORRESPONDENCE ADDRESS:
(A) ADDRESSEE: LERNER, DAVID, LITTENBERG, KRUMHOLZ & MENTILIK
(B) STREET: 600 South, Avenue West
(C) CITY: Westfield
(D) STATE: New Jersey
(E) COUNTRY: USA
(F) ZIP: 07090

(v) COMPUTER READABLE FORM:
(A) MEDIUM TYPE: Floppy disk
(B) COMPUTER: IBM PC compatible
(C) OPERATING SYSTEM: PC-DOS/MS-DOS
(D) SOFTWARE: PatentIn Release #1.0, Version #1.30

(vi) CURRENT APPLICATION DATA:
(A) APPLICATION NUMBER: US 07/662,764
(B) FILING DATE: 28-FEB-1991
(C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: US 07/201,358
(B) FILING DATE: 26-MAY-1988

(vii) PRIOR APPLICATION DATA:
(A) APPLICATION NUMBER: US 06/770,390
(B) FILING DATE: 28-AUG-1985

(viii) ATTORNEY/AGENT INFORMATION:
(A) NAME: Foley, Shawn P.
(B) REGISTRATION NUMBER: 33,071
(C) REFERENCE/DOCKET NUMBER: ICTECH/0002

(ix) TELECOMMUNICATION INFORMATION:
(A) TELEPHONE: 908-654-5000
(B) TELEFAX: 908-654-7866

(2) INFORMATION FOR SEQ ID NO:1:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 24 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: other nucleic acid
(A) DESCRIPTION: /desc = "oligonucleotide"

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

GATCCTTNNN NNNNNNNNNN NNAA 24

(2) INFORMATION FOR SEQ ID NO:2:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

CTTACCGAGC GGACTGGTAA A 21

(2) INFORMATION FOR SEQ ID NO:3:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CTTATGCAAG ACTCGATACA A 21

(2) INFORMATION FOR SEQ ID NO:4:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

CTTGCGGGGT CAGAGGGCGA A 21

(2) INFORMATION FOR SEQ ID NO:5:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: liricar

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

CTTCAGATAT TTCCGAAGCA A 21

(2) INFORMATION FOR SEQ ID NO:6:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

CTTAACATCC TCCAACGGCA A 21

(2) INFORMATION FOR SEQ ID NO:7:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid

(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

CTTCCATCGC TGAAACTCAA A 21

(2) INFORMATION FOR SEQ ID NO:8:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

CTTACACCGA GGGCGCTCCA A 21

(2) INFORMATION FOR SEQ ID NO:9:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

CTTCTAGAAT TCGTGGGCAA A 21

(2) INFORMATION FOR SEQ ID NO:10:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

CTTAGCGTGC TCGACAGGCA A 21

(2) INFORMATION FOR SEQ ID NO:11:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

CTTCAAGACA AAGTACATCA A 21

(2) INFORMATION FOR SEQ ID NO:12:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

CTTGAAGTAT ATCAAGCAGA A 21

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

CTTGTTTTCC TTACTCCCGA A 21

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

CTTCTATACA TAACCAACAA A 21

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 21 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

CTTGACGCGG ATATAGGAAA A 21

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

TCTTTACCAG TCCGCTCGGT AAGATCCTCA TGAGGATCTT ACC GAG CGG ACT GGT 55
Thr Glu Arg Thr Gly
1 5

AAA GA 60
Lys

(2) INFORMATION FOR SEQ ID NO:17:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:
Thr Glu Arg Thr Gly Lys
1 5

(2) INFORMATION FOR SEQ ID NO:18:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

TCTTGTATCG AGTCTTGCAT AAGATCCTCA TGAGGATCTT ATG CAA GAC TCG ATA 55
Met Gln Asp Ser Ile
1 5

CAA GA 60
Gln

(2) INFORMATION FOR SEQ ID NO:19:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

Met Gln Asp Ser Ile Gln
1 5

(2) INFORMATION FOR SEQ ID NO:20:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

TCTTCGCCCT CTGACCCCGC AAGATCCTCA TGAGGATCTT GCG GGG TCA GAG GGC 55
Ala Gly Ser Glu Gly

1 5

GAA GA 60
Glu

(2) INFORMATION FOR SEQ ID NO:21:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

Ala Gly Ser Glu Gly Glu
1 5

(2) INFORMATION FOR SEQ ID NO:22:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

TCTTGCTTCG GAAATATCTG AAGATCCTCA TGAGGATCTT CAG ATA TTT CCG AAG 55
Gln Ile Phe Pro Lys
1 5

CAA GA 60
Gln

(2) INFORMATION FOR SEQ ID NO:23:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

Gln Ile Phe Pro Lys Gln
1 5

(2) INFORMATION FOR SEQ ID NO:24:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

TCTTGCCGTT GGAGGATGTT AAGATCCTCA TGAGGATCTT AAC ATC CTC CAA CGG 55
Asn Ile Leu Gln Arg
1 5

CAA GA 60
Gln

(2) INFORMATION FOR SEQ ID NO:25:

(i) SEQUENCE-CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

Asn Ile Leu Gln Arg Gln
1 5

(2) INFORMATION FOR SEQ ID NO:26:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

TCTTTGAGTT TCAGCGATGG AAGATCCTCA TGAGGATCTT CCA TCG CTG AAA CTC 55
Pro Ser Leu Lys Leu
1 5

AAA GA 60
Lys

(2) INFORMATION FOR SEQ ID NO:27:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

Pro Ser Leu Lys Leu Lys
1 5

(2) INFORMATION FOR SEQ ID NO:28:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 40..57

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

TCTTGGAGCG CCCTCGGTGT AAGATCCTCT GAGGATCTT ACA CCG AGG GCG CTC 54
Thr Pro Arg Ala Leu
1 5

CAA GA 59
Gln

(2) INFORMATION FOR SEQ ID NO:29:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

Thr Pro Arg Ala Leu Gln
1 5

(2) INFORMATION FOR SEQ ID NO:30:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

TCTTTGCCCA CGAATTCTAG AAGATCCTCA TGAGGATCTT CTA GAA TTC GTG GGC 55
Leu Glu Phe Val Gly
1 5

AAA GA 60
Lys

(2) INFORMATION FOR SEQ ID NO:31:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

Leu Glu Phe Val Gly Lys
1 5

(2) INFORMATION FOR SEQ ID NO:32:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENcE DESCRIPTION: SEQ ID NO:32:

TCTTGCCTGT CGAGCACGCT AAGATCCTCA TGAGGATCTT AGC GTG CTC GAC AGG 55
Ser Val Leu Asp Arg
1 5

CAA GA 60
Gln

(2) INFORMATION FOR SEQ ID NO:33:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids

(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

Ser Val Leu Asp Arg Gln
1 5

(2) INFORMATION FOR SEQ ID NO:34:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

TCTTGATGTA CTTTGTCTTG AAGATCCTCA TGAGGATCTT CAA GAC AAA GTA CAT 55
Gln Asp Lys Val His
1 5

CAA GA 60
Gln

(2) INFORMATION FOR SEQ ID NO:35:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

Gln Asp Lys Val His Gln
1 5

(2) INFORMATION FOR SEQ ID NO:36:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

TCTTCTGCTT GATATACTTC AAGATCCTCA TGAGGATCTT GAA GTA TAT CAA GCA 55
Glu Val Tyr Gln Ala
1 5

GAA GA 60
Glu

(2) INFORMATION FOR SEQ ID NO:37:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

Glu Val Tyr Gln Ala Glu
1 5

(2) INFORMATION FOR SEQ ID NO:38:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:38;

TCTTCGGGAG TAAGGAAAAC AAGATCCTCA TGAGGATCTT GTT TTC CTT ACT CCC 55
Val Phe Leu Thr Pro
1 5

GAA GA 60
Glu

(2) INFORMATION FOR SEQ ID NO:39:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

Val Phe Leu Thr Pro Glu
1 5

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 60 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

TCTTTGTTGG TTATGTATAG AAGATCCTCA TGAGGATCTT CTA TAC ATA ACC AAC 55
Leu Tyr Ile Thr Asn

1 5

AAA GA 60
Lys

(2) INFORMATION FOR SEQ ID NO:41:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:41:

Leu Tyr Ile Thr Asn Lys
1 5

(2) INFORMATION FOR SEQ ID NO:42:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 59 base pairs
(B) TYPE: nucleic acid
(C) STRANDEDNESS: single
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(ix) FEATURE:
(A) NAME/KEY: CDS
(B) LOCATION: 41..58

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:42:

TCTTTTCCTA TATCCGCGTC AAGATCCTCA TGAGGATCTT GAC GCG GAT ATA GGA 55
Asp Ala Asp Ile Gly
1 5

AAG A 59
Lys

(2) INFORMATION FOR SEQ ID NO:43:

(i) SEQUENCE CHARACTERISTICS:

(A) LENGTH: 6 amino acids
(B) TYPE: amino acid
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: protein

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:43:

Asp Ala Asp Ile Gly Lys
1 5

(2) INFORMATION FOR SEQ ID NO:44:

(i) SEQUENCE CHARACTERISTICS:
(A) LENGTH: 16 amino acids
(B) TYPE: amino acid
(C) STRANDEDNESS:
(D) TOPOLOGY: linear
(ii) MOLECULE TYPE: peptide

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:44:

Asp Asp Glu Val Asp Val Asp Gly Thr Val Glu Glu Asp Leu Gly Tyr
1 5 10 15

IN THE CLAIMS:

1. A discrete population of oligonucleotides of random sequence wherein:

each member of the population comprises between 1 and about 50 tandem sequences of from about 4 to about 12 nucleotide triplets, **the order and selection of said triplets being random,**{

each member of the population has the same number of tandem repeating sequences of the same length,}

each of the tandem sequences encodes {for} a corresponding peptide sequence of **one of from** about 4 to about 12 L-amino acid residues, and

{the population encodes for} **wherein the sum of said corresponding peptide sequences represents** at least about 10% of all **possible corresponding** peptide sequences of the selected length.

2. The oligonucleotide population of claim 1 wherein each member of the population comprises a single copy of the sequence of nucleotide triplets.

3. The oligonucleotide population of claim {1} **1** wherein each sequence comprises **a single length of** from {5} **4** to 7 nucleotide triplets.

4. The oligonucleotide population of claim {2} **1** wherein the population is generated by shearing of mammalian genetic material.

5. The oligonucleotide population of claim 1 wherein the population is chemically synthesized from the component {nucleic acids.} **nucleotides.**

{6. A population of peptides, wherein:
each member of the population comprises from 1 to about 50 tandem peptide sequences of about 4 to about 12 L-amino acid residues,
each member of the population has the same number of tandem sequences of the same length, and
the population contains at least about 10% of all possible peptide sequences of the selected length.

7. The peptide:/,population of claim 6 wherein each member of the population comprises a single copy of the peptide sequence.

8. The peptide population of claim 6 wherein each sequence comprises from 5 to 7 L-amino acid residues.

9. The peptide population of claim 7 wherein the population is generated by shearing of proteins.

10. The population of claim 6 wherein the population is chemically synthesized from the component L-amino acids.

11. A} **6.** **A discrete** population of vectors{,} comprising:
substantially identical autonomously replicating nucleic acidsequences wherein at least a portion of each **nucleic acid** sequence is a structural gene, and

oligonucleotide inserts {comprising a population }**consisting essentially of a population of oligonucleotide inserts wherein each insert consists essentially of one** of from 1 to about 50 tandem {units of }**sequences, each of which is from** about 4 to about 12 nucleotide triplets, **the selection and order of said triplets being random, and** wherein each oligonucleotide insert has the same number of tandem {units of} **sequences, and each sequence has** the same length, **and**

wherein each member of the population of oligonucleotide inserts is recombinantly inserted into the structural gene of {one} **each** of the replicating sequences{, a significant proportion of the vectors are} **to form a recombinant structural gene, and**

wherein a significant portion of the vectors is capable of expressing {their} **the** recombinant structural {genes} **gene** when transferred into appropriate host cells, and **wherein** expression of the recombinant structural genes yields polypeptides {comprising from 1}**, each comprising a corresponding peptide sequence encoded by the oligonucleotide insert and each comprising one length of from 1** to about 50 tandem peptide sequences **each** of **from** about 4 to about 12 L-amino acid residues encoded by {their} **the** respective oligonucleotide inserts.

{12} 7. The vector population of claim {11} **6** wherein each member of the oligonucleotide insert population comprises a single copy of the sequence of nucleotide triplets.

{13} **8**. The vector population of claim {11,} **6** wherein each {unit} **tandem sequence** comprises **a single length of** from {5} **4** to 7 nucleic acid triplets.

{14} **9**. The peptide population produced by the vector population of claim {11} **6**.

{15} **10**. The peptide population of claim {14} **9**, wherein each repeating sequence comprises **a single length of** from {5} **4** to 7 nucleic acid triplets.

{16} **11**. The vector population of claim {11} **6** wherein the replicating sequence is a plasmid.

{17} **12**. The vector population of claim {16} **11** wherein the plasmid is pBR322.

13. The vector population of claim 11 wherein the plasmid is pUC8.

14 {18}. The vector population of claim {11} **6** wherein the replicating sequence is {a virus} **viral**.

{19} **15**. The vector population of claim {18} **14** wherein the {virus} **viral replication sequence** is lambda-gt 11.

{20} **16**. The vector population of claim {18,} **14** wherein the {virus is a strain} **viral replicating sequence is a derivative** of vaccinia.

{21} **17**. The vector population of claim {11} **6** wherein the replicating sequence comprises a filamentous bacteriophage.

{22} **18**. The vector population of claim {21} **17** wherein the filamentous {bacteriophage is pUC8.} **bacteriophase is λ.**

{23. A method of modifying a vector to create a modified vector possessing an epitope on its outside surface and of identifying the modified vector, the method} **19. the discrete oligonucleotide population of claim 1 wherein each of said encoded corresponding peptide sequences is capable of forming a binding air with an antibody that has not been elicited by immunization with said peptide sequence or said peptide sequence in conjugated form, said antibody being selected from the group consisting of all antibodies produced by lymphoid-derived antibody producing cells.**

20. The vector population of claim 1 wherein each of the generated polypeptides is capable of forming a binding pair with an antibody that has not been elicited by immunization with said peptide or said peptide in conjugated form, said antibody being selected from a group consisting of all antibodies produced by lymphoid-derived antibody producing cells.

21. A method of producing a discrete population of random epitopic peptide sequences, wherein said epitopic seqeunces are each accessible to antibody recognition, comprising the steps of:

{isolating a plurality of an appropriate vector comprising an autonomously replicating DNA element,} **inserting oligonucleotides of random amino acid coding sequence of a population of vector nucleic acid molecules within a structural gene to produce a recombinant vector such that said oligonucleotides express random epitopic peptide sequences in a position where said epitopic peptide sequences are accessible to antibody recognition**

{cleaving the circular DNA at random with respect to nucleotide sequence, producing a population of linear DNA molecules comprising circular permutations of the same nucleotide sequence, }**whereby said discrete population of random epitopic peptide sequences comprises substantially every epitopic peptide sequence.**

{joining a unique oligonucleotide sequence to the ends of the linear DNA molecules, the unique oligonucleotide sequence not otherwise existing in the DNA element and comprising at least a portion of a structural gene of a foreign organism,

rejoining the ends to form circular double stranded DNA molecules having the oligonucleotide of unique sequence inserted at random with respect to the nucleotide sequence of each circular DNA molecule,

transferring the circular DNA having the unique insert sequence to a host organism under conditions permitting replication of the DNA,

screening the progeny of the circular DNA having a unique insert sequence with a monoclonal or polyclonal antibody that recognizes the foreign structural gene, the progeny bearing the insert in a non-essential region of the DNA and expressing the insert in such a manner that its product is recognized by the antibody.

24. A heterogeneous populations of antibodies, comprising antibodies capable of binding to substantially every member of the peptide population of claim 6.

25} **22**. A method of producing a {heterogeneous} **discrete** population of antibodies, the method comprising the steps of:

harvesting **mammalian** lymph cells {from a ma that has} **wherein said lymph cells have** not been antigenically stimulated {with a parti} **witha particular** antigen{,}**;**

fusing {the} **said** lymph cells with myeloma cells to produce hybridoma cells{,}**;** and

culturing individual hybridoma {cell} **cells** lines, {the} **said** cell lines **being** capable of producing {antigens} **antibodies** that are capable of recognizing a broad range of antigens

{.} **whereby said discrete population of antibodies comprising antibodies which together recognize substantially all epitopic peptide sequences.**

{26} **23**. The method of claim {25, wherein the mammal is raised aseptically until the lymph cells are harvested.} **22 wherein said lymph cells are prepared from mammals raised in an antiseptic environment.**

{27. The method claim 25, wherein the} **24.**

The method of claim 22 wherein said lymph cells are harvested from a {fetal

mammal or from a neonatal mammal not yet capable of responding to antigenic stimulation.} **mammal which is a neonate and a fetus.**

{28. A }**25. A discrete** population of binding pairs comprising:

a **discrete** population of peptide sequences of the same length, the length being **selected from lengths of from** about 4 to about 12 {L-}amino acid residues{, the} **and wherein said** population {comprising at least 10% of} **is sufficiently large to comprise substantially** all peptide sequences of {the} **said** selected length{, and};

a {heterogeneous} **discrete** population of antibodies comprising antibodies capable of binding to substantially every member of {the oligopeptide population,

}**said population of peptide sequences,**

wherein within said population of binding pairs substantially every member of the peptide population {being bounded to its corresponding} **is bound with an** antibody.

{29} **26.** A matrix comprising **the discrete population of binding pairs of claim 25.**{:}

{a population of peptide sequences of the same length, the length being about 4 to about 12 L-amino acid residues, the population comprising at least 10% of all peptide sequences of the selected length, and

a heterogeneous population of antibodies comprising antibodies capable of binding to substantially every member of the oligopeptide population.

30. The matrix of claim 29, wherein each of the peptide sequences is immobilized on an appropriate substrate and the immobilized peptide sequences are contacted with the antibodies.

31. The matrix of claim 30, wherein each of the antibodies is labeled with an appropriate label that does not interfere substantially with binding and provides a means for identifying binding pairs.

32. The matrix of claim 29 wherein each of the antibodies is immobilized on an appropriate substrate and the immobilized antibodies are contacted with the peptide sequences.

33. The matrix of claim 32, wherein each of the peptide sequences is labeled with an appropriate label that does not interfere substantially with binding and provides a means for identifying binding pairs.

34. The matrix of claim 33 wherein each of the peptide sequences is located on the surf a fusion protein or modified vector, the protein tor itself comprising the label.

35. The matrix of claim 29 wherein each of the peptide sequences is contacted with each of the antibodies until at least one peptide sequence-antibody binding pair is identified.

36} **27**. A submatrix comprising:

a {population of peptide sequences of the same length, the length being about 4 to about 12 L-amino acid residues, the population comprising a significant proportion of those peptide sequence of the selected length having sufficient conformational similarity with the antibody binding sites of a test species that an antibody capable of binding to an antibody binding site of the test species is also capable of binding to a member of the peptide population,

a heterogeneous population of antibodies comprising antibodies capable of binding to substantially every member of the oligopeptide population.

37. The submatrix of claim 36 wherein each of the peptide sequences is contacted with each of the antibodies until at least one individual antibody-peptide sequence binding pair is identified.

38. The submatrix of claim 36 wherein each of the peptide sequences is immobilized on an appropriate substrate and the immobilized peptide sequences are contacted with the antibodies.

39. The submatrix of claim 36 wherein each of the antibodies is immobilized on an appropriate substrate and the immobilized antibodies are contacted with the peptide sequences.

40. The submatrix of claim 36 wherein the test species is a virus or bacteriophage.

41. The submatrix of claim 36 wherein the test species is selected from the group of enzymes, proteins, and polypeptides.

42. The submatrix of claim 36 wherein the test species is selected from the group of non-peptide drugs and non-peptide bioactive substances.

43. A method for constructing a matrix comprising:

obtaining a population of peptide sequences having about 4 to about 12 L-amino acid residues, each member of the population having the same length, the population comprising at least 10% of all peptide sequences of the predetermined length,

obtaining a heterogeneous population of antibodies, comprising antibodies capable of binding to substantially every member of the peptide sequence population, and

contacting the antibodies with the peptide sequences for a sufficient amount of time and under appropriate conditions so that at least one peptide sequence-antibody binding pair is created.

44. The method o claim 43, further comprising the step of:

labeling the antibodies, the peptide sequences, or both with an appropriate label that does not interfere substantially with binding and provides a means for identifying any binding pairs.

45. The method of claim 43, wherein each of the peptide sequences is purified, each of the antibodies is purified, and each of the peptide sequences is contacted with each of the antibodies until at least one peptide sequence-antibody binding pair is identified.

46. The method of claim 45, wherein each of the peptide sequences is contacted individually with each of the antibodies until at least one peptide sequence-antibody binding pair is identified.

47. The method of claim 43, wherein each of the peptide sequences is immobilized on an appropriate substrate and the immobilized peptide sequences are contacted with the antibodies.

48. The method of claim 47, wherein each of the antibodies is labeled with an appropriate label that does not interfere substantially with binding and provides a means for identifying binding pairs.

49. The method of claim 43, wherein each of the antibodies is immobilized on an appropriate substrate and the immobilized antibodies are contacted with the peptide sequences.

50. The method of claim 49, wherein each of the peptide sequences is labeled with an appropriate label that does not interfere substantially with binding but provides a means for identifying binding pairs.

51. The method of claim 50 wherein each of the peptide sequences is located on the surface of a fusion protein or modified vector, the protein or vector itself comprising the label.

52. The method of claim 43 wherein each of the peptide sequences is translated from a genetically engineered vector as a portion of a larger fusion polypeptide.

53. The method of claim 52 wherein each peptide sequence is excised from its parent polypeptide.

54. A method for determining immunological and/or genotypic properties of a test species, wherein the test species is an antibody, virus, bacteriophage, enzyme, protein, polypeptide, non-peptide drug, or non-peptide bioactive substance, the method comprising the steps of:

constructing a matrix comprising a population of peptide sequences of the same length, the length being about 4 to about 12 L-amino acid residues, the population comprising at least 10% of all peptide sequences of the selected length; and a heterogeneous pop of antibodies comprising antibodies capable of binding to substantially every member of the peptide sequence population;

contacting the antibodies with the peptide sequences, for a sufficient amount of time and under appropriate conditions so that at least one peptide sequence-antibody binding pair is created,

contacting the test species with the matrix,

observing where the test species disturbs the binding pairs, and identifying the peptide sequence or the antibody at the site or sites where binding is disturbed.

55. A method of identifying a specific peptide sequence that has sufficient conformational similarity to an antibody recognition site on a test species that an antibody capable of recognizing and binding to the recognition site may also be capable of binding to the peptide sequence, the method comprising the steps of:

contacting the test species with a matrix,

observing where the test species disturbs the binding pairs, and

identifying a peptide sequence comprising a binding pair disturbed by the presence of the test species.

56. The method of claim 55, wherein the matrix is a peptide immobilized matrix wherein each immobilized peptide sequence forms a binding pair with a

corresponding antibody, and the peptide sequence immobilized at a site where binding is disturbed is identified.

57. The method of claim 55, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide sequence displaced by the presence of the test species is identified.

58. A method of developing a vaccine against a disease producing agent, the method comprising the steps of:
contacting the disease producing agent with a matrix, observing where the disease producing agent disturbs the binding pairs,
identifying the peptide sequence comprising a binding pair disturbed by the presence of the disease producing agent, and
constructing an antigen comprising the peptide sequence.

59. The method of claim 58, wherein the matrix is a peptide immobilized matrix wherein each immobilized peptide sequence forms a binding pair with a corresponding antibody, and the peptide sequence immobilized at a site where binding is disturbed is identified.

60. The method of claim 58, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide sequence displaced by the presence of the disease associated substance is identified.

61. A method of characterizing a recombinant gene product of a gene expression library, the method comprising the steps of:

contacting the recombinant gene product with a matrix,

observing where the gene product disturbs the binding pairs, and

identifying the peptide sequence comprising a binding pair disturbed by the presence of the recombinant gene product.

62. The method of claim 61, wherein the matrix is a peptide immobilized matrix wherein each immobilized peptide sequence forms a binding pair with a corresponding antibody, and the peptide sequence immobilized at a site where binding is disturbed is identified.

63. The method of claim 61, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide sequence displaced by the presence of the recombinant gene product is identified.

64. A method of locating, in a genome, the gene encoding for a protein, enzymes, or peptide, the method comprising:

contacting the protein, enzyme, or peptide with a matrix,

observing where the protein disturbs the binding pairs,

identifying the recombinant cell line that produced a peptide sequence comprising a binding pair disturbed by the presence of the protein, enzyme, or peptide, and

using the nucleotide sequence of the oligonucleotide insert encoding for the peptide sequence as a DNA probe to locate the gene encoding for the protein.

65. The method of claim 64, wherein the matrix is a peptide immobilized matrix wherein each immobilized peptide sequence forms a binding pair with a corresponding antibody, and the peptide sequence immobilized at a site where binding is disturbed is identified.

66. The method of claim 64, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide sequence displaced by the presence of the protein, enzyme, or peptide is identified.

67. A method of determining a peptide sequence recognized by a first antibody, the method comprising the steps of:
contacting the first antibody with a matrix,
observing where the first antibody binds to a matrix-associated peptide sequence, and identifying the peptide sequence.

68. The method of claim 67, wherein the matrix is a peptide immobilized matrix and the peptide sequence immobilized at a site where binding is disturbed is identified.

69. The method of claim 67, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide displaced by the presence of the first antibody is identified.

70. A method of determining the nucleotide sequence that encodes for a peptide sequence recognized by a first antibody, the method comprising the steps of:
contacting the first antibody with a matrix,

observing where the first antibody binds to a matrix-associated peptide sequence,

identifying the genetically recombinant cell line that produced the peptide sequence, and

determining the sequence of the oligonucleotide encoding for the peptide sequence inserted in the vector transferred into the cell line.

71. The method of claim 70, wherein the matrix is a peptide immobilized matrix and the peptide sequence immobilized at a site where binding is disturbed is identified.

72. The method of claim 70, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide sequence displaced by the presence of the first antibody is identified.

73. A method for treating a human patient suffering from an autoimmune disease wherein antibodies produced by the patient recognize and impair the functioning of the patient's own cells, the method comprising the steps of:

isolating antibodies produced by the patient that recognize the patient's own cells,

contacting the antibodies with a matrix,

observing where the antibodies disturb the binding pairs,

identifying the peptide sequence comprising a binding pair disturbed by the presence of the antibodies, and

administering to the patient an effective amount of the peptide sequence to competitively inhibit in vivo the antibodies from binding to the patient's own cells and thereby to improve the condition of the patient.

74. The method of claim 73, wherein the matrix is a peptide immobilized matrix and the peptide sequence immobilized at a site where binding is disturbed is identified.

75. The method of claim 73, wherein the matrix is an antibody immobilized matrix wherein each immobilized antibody forms a binding pair with a corresponding peptide sequence, and a peptide sequence displaced by the presence of the antibodies produced by the patient is identified.

76. A method of identifying and selecting an antibody that reacts with a test species, the method comprising the steps of:
contacting the test species with an antibody immobilized matrix,
observing where the test species binds to an immobilized antibody, and
identifying the antibody immobilized at a site where binding occurs.

77. A method of testing for the presence of a test species, the method comprising the steps of:
contacting the test species with an antibody-immobilized matrix,
observing where the test species binds to an immobilized antibody,
identifying the antibody immobilized at a site where binding occurs,
culturing a hybridoma cell line from which the identified antibody was derived to provide a source of the identified antibody, and
using the identified antibody in an immunnoassay to test for the presence of the test species.

78. A diagnostic test comprising the steps of:
contacting a disease associated substance with an antibody-immobilized matrix,

observing where the disease associated substance binds to an immobilized antibody,
identifying the antibody immobilized at a site where the binding occurs,
culturing the hybridoma cell line from which the identified antibody was derived to provide a source of the identified antibody, and
contacting the antibody with an appropriate sample from a patient to test for the presence of the disease associated substance.

79. a diagnostic test kit comprising an antibody that recognizes an epitope on a disease associated substance, wherein the antibody is prepared by:
contacting the disease associated substance with an antibody-immobilized matrix,
observing where the disease associated substance binds to an immobilized antibody,
identifying the antibody immobilized at a site where binding occurs, and
culturing the hybridoma cell line from which the identified antibody was derived to provide a source of the indentified antibody.

80. A method for targeting a drug in a human patient to a specific class of malignant cells, the method comprising the steps of:
isolating a first sample of malignant cells from the patient and a second sample of healthy cells from the patient,
contacting the first cell sample with an antibody immobilized matrix,
observing where the first cell sample binds to the matrix and identifying the immobilized antibodies that bind to the first cell sample,
screening the identified antibodies for reactivity with a second cell sample and
selecting those antibodies capable of binding to members of the first cell sample but incapable of binding to members of the second cell sample,

culturing at least one of the hybridoma cell lines from which the selected antibodies were derived to provide a source of the selected antibodies,

linking the drug molecules to a population of antibodies comprising the selected antibodies, and

administering a malignant-cell-growth-affecting amount of the drug-linked antibodies to the patient.

METHOD AND MEANS FOR SORTING AND IDENTIFYING BIOLOGICAL INFORMATION} **discrete population of random peptide sequences wherein the population of peptide sequences comprises a subpopulation of peptide sequences, which subpopulation comprises a significant proportion of peptide sequences of the selected length having sufficient conformational similarity with an antibody binding site of a test species such that an antibody capable of binding to the antibody binding site of the test species is also capable of binding to a member of the discrete population of peptide sequences; and**

a heterogeneous discrete population of antibodies comprising antibodies capable of binding to substantially every member of the subpopulation of peptide sequences.

28. A method for creating the matrix of claim 27 comprising:

obtaining a discrete population of random peptide sequences wherein said discrete population of peptide sequences comprises at substantially all possible peptide sequences;

obtaining a discrete heterogeneous population of antibodies, wherein said population of antibodies comprising antibodies capable of binding to substantially every member of the discrete population of peptide sequences; and

contacting said discrete population of peptide sequences with said discrete heterogeneous population of antibodies for a sufficient amount of time under appropriate conditions so that at east one peptide sequence-antibody binding, pair is created.

Method and Means for Sorting and Identifying Biological Information

ABSTRACT OF THE DISCLOSURE

In one aspect the invention discloses a matrix comprising a **discrete** population of {peptide sequences} **random oligopeptides** of the same length, the length being **selected from** about 4 to about 12 L-amino acid residues, the population comprising at least 10% of all {peptide} **amino acid** sequences of the selected length; and a heterogeneous population of antibodies comprising antibodies capable of binding to substantially every member of the oligopeptide population.